Plant Genomics and Proteomics

Plant Genomics and Proteomics

Christopher A. Cullis

WILEY-LISS

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2004 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:
Cullis, Christopher A., 1945–
Plant genomics and proteomics / Christopher A. Cullis.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-37314-1
1. Plant genomes. 2. Plant proteomics. I. Title.
QK981.C85 2004
572.8′62—dc21
2003013088

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments, vii

Introduction, ix

1 The Structure of Plant Genomes, 1

2 The Basic Toolbox—Acquiring Functional Genomic Data, 23

3 Sequencing Strategies, 47

4 Gene Discovery, 69

5 Control of Gene Expression, 89

6 Functional Genomics, 107

7 Interactions with the External Environment, 131

8 Identification and Manipulation of Complex Traits, 147

9 Bioinformatics, 167

10 Bioethical Concerns and the Future of Plant Genomics, 189

Afterword, 201

Index, 203

Acknowledgments

This book would not have been possible without the contributions of two individuals. First, I would like to thank my wife Margaret, whose efforts in reading the drafts and suggesting clarifications were invaluable. Any obscure or erroneous passages are certainly not her responsibility; she probably just could not get me to change my mind. Second, I would like to thank my son Oliver, with whom I shared the first attempts at writing a book and who contributed with comments on the clarity of early drafts.

Introduction

What possible rationale is there for developing a genomics text that is focused on only the plant kingdom? Clearly, there are major differences between plants and animals in many of their fundamental characteristics. Plants are usually unable to move, they can be extremely long lived, and they are generally autotrophic and so need only minerals, light, water, and air to grow. Thus the genome must encode the enzymes that support the whole range of necessary metabolic processes including photosynthesis, respiration, intermediary metabolism, mineral acquisition, and the synthesis of fatty acids, lipids, amino acids, nucleotides, and cofactors, many of which are acquired by animals through their diet. At a technological level genomics studies, which take a global view of the genomic information and how it is used to define the form and function of an organism, have a common thread that can be applied to almost any system. However, plants have processes of particular interest and pose specific problems that cannot be investigated in any one simple model and often even need to be investigated in a particular plant species. Plant genomics builds on centuries of observations and experiments for many plant processes. Because of this history, much of the experimental detail and observations span very diverse plant material, rather than all being available in a convenient single model organism. Thus algae may be appropriate models for photosynthesis and provide useful pointers as to which genes are involved but, conversely, cannot be useful for understanding, for example, how stresses in the roots might affect the same photosynthetic processes in a plant growing under drought or saline conditions. The genomics approaches to plant biology will result in an enhanced knowledge of gene structure, function, and variability in plants. The application of this new knowledge will lead to new methods of improving crop production, which are necessary to meet the challenge of sustaining our food supply in the future.

One of the particularly relevant differences, for this text, between plants and other groups of organisms is the large range of nuclear DNA contents

(genome sizes) that occur in the plant kingdom, even between closely related species. Therefore, it is harder to define the nature of a typical plant genome because the contribution of additional DNA may have phenotypic effects independent of the actual sequences of DNA present, for example, the role of nuclear DNA content in the annual versus perennial life cycle. An added complication is that rounds of polyploidization followed by a restructuring of a polyploid genome have frequently occurred during evolution. The restructuring of the genome has usually resulted in a loss of some of the additional DNA derived from the original polyploid event. Therefore, the detailed characterization of a number of plant genomes, rather than a single model or small number of models, will be important in developing an understanding of the functional and evolutionary constraints on genome size in plants. Despite this enormous variation in DNA content per cell, it is generally accepted that most plants have about the same number of genes and a similar genetic blueprint controlling growth and development.

As indicated in the opening paragraph, the wealth of data for many processes, such as cell wall synthesis, photosynthesis and disease resistance, has been generated by investigating the most amenable systems for understanding that particular process. However, many of these models are not well characterized in other respects and have relatively few genomics resources, such as sequence data and extensive mutant collections, associated with them. Therefore, the information derived from each of these systems will have to be confirmed in a well characterized model plant to understand the molecular integration and coordination of development for many of the intertwined pathways. This may not be possible in the best-characterized systems of each of the individual elements. *Zinnia* provides an excellent model to study the differentiation of tracheary elements because isolated mesophyll cells can be synchronously induced to form these elements *in vitro*. Therefore, this synchrony permits the establishment and chronology of the molecular and biochemical events associated with the differentiation of the cells to a specific fate and the identification of the genes involved in the differentiation of xylem. However, *Zinnia* does not have the experimental infrastructure to allow extensive genomic investigations into other important processes. Therefore, the detailed knowledge acquired would need to be integrated in another more fully described model plant, although the knowledge would have been difficult to identify without resource to this specialized experimental system. Therefore, the accumulation of genomic information will be necessary across the plant kingdom, with an integrated synthesis perhaps finally occurring only in a few model species. The relevant approaches will include the development of detailed molecular descriptions of the myriad of plant pathways for many plant species in order to unravel the secrets of how plants grow, develop, reproduce, and interact with their environments.

The publication of the *Arabidopsis* and rice genomic sequences has

facilitated the comparison between plants and animals at the sequence level. Not surprisingly, perhaps, the initial comparisons have shown that some processes, such as transport across membranes and DNA recombination and repair processes, appear to be conserved across the kingdoms whereas others are greatly diverged. Many novel genes have been found in the plant genomes so far characterized, which was expected considering the wide range of functions that occur in plants but are absent from animals and microbes.

The easy access to plant genome sequences and all of the other genomics tools, such as tagged mutant collections, microarrays, and proteomics techniques, has fundamentally changed the way in which plant science can be done. Old problems that appeared to be intractable can now be tackled with renewed vigor and enthusiasm. One example is the Floral Genome Project (http://128.118.180.140/fgp/home.html) tackling what Darwin referred to as "The abominable mystery," namely, the origin of flowering plants, that has gone unanswered for more than a century. More than just answering this question, though, the origin and diversification of the flower is a fundamental problem in plant biology. The structure of flowers has major evolutionary and economic impacts because of their importance in plant reproduction and agriculture.

The two different regions of the plant, the aerial portions (stems, leaves, and flowers) and the below-ground portions (roots), have received very different treatment as far as experimental investigations are concerned. The above-ground regions of the plant have clearly been more amenable to visual description and biochemical characterization. This is partly due to the difficulty in studying the roots. Not only are they normally in a nonsterile environment, beset with many microorganisms both beneficial and harmful, but they are also difficult to separate from the physical medium of the soil. As genomic tools continue to be developed it will become easier to delineate the contribution and characteristics of the associated microorganisms and the plant roots and so understand the interaction of the roots and the microenvironment in the soil. Of particular interest is the understanding of the beneficial interactions between the plant roots and microorganisms such as rhizobia and mycorrhizae, in contrast to the destructive interactions between the roots and pathogens.

The interface between the plant and pathogens is also important with respect to the aerial portions of a plant. The combination of an increased understanding of the pathogen's genome, as well as the responses that occur in both the pathogen and the host on infection, will open up new methods for controlling diseases in crops. The detailed understanding of the interplay between the plant and the pathogen should also enable the development and incorporation of more durable resistances to many of the destructive plant diseases, resulting in an increased security of the food supply worldwide. Therefore, these new interventions, supported by information from

genomics studies, will be important both for increasing yield and for reducing environmental hazards that may be associated with the current agronomic use of available fungicides and insecticides.

Light, as well as being the primary energy source for plants, also acts as a regulator of many developmental processes. Chlorophyll synthesis and the induction of many nucleus- and chloroplast-encoded genes are affected by both light quality and quantity. In this respect the close coupling of the nuclear and chloroplast genomes is another unique plant process. Many of the biochemical reactions of light responses have already been well documented, but the ability to recognize the genes that have been transferred from the organellar genomes to the nucleus may also shed light both on the coordinated control of these responses and on the evolutionary history, pressures, and constraints. Again, the input from the characterization of the genomes of algae and other microorganisms will greatly facilitate all such studies.

The synthesis of cell walls and their subsequent modification are clearly important processes in higher plants. The initial annotation of the *Arabidopsis* genome identified more than 420 genes that could tentatively be assigned roles in the pathways responsible for the synthesis and modification of cell wall polymers. The fact that many of these genes belong to families of structurally related enzymes is also an indication of the apparent gene redundancy in the plant genome. However, as will be discussed in this work, whether this redundancy is real, in the sense that one member of the family can effectively substitute for any of the other members, or whether this is only an apparent redundancy and the various genes reflect differences in substrate specificity or developmental stage at which they function, is still to be determined.

Plants synthesize a dazzling array of secondary metabolites. More than a hundred thousand of these are made across all species. The exact nature and function of most of these metabolites still await understanding. The combination of information from sequencing, expression profiling, and metabolic profiling will help to define the relationship between the genes involved, their expression, and the synthesis of these metabolites. The understanding of which member of a gene family is expressed in a particular tissue, and the specific reaction in which it is involved, will also shed light on the level of redundancy of gene functions for the synthesis of many of these compounds.

Many of the processes that are known to regulate or control development in animals including the modulation of chromatin structure, the cascades of transcription factors, and cell-to-cell communications, will also be expected to regulate plant development. However, the initial analysis of the *Arabidopsis* genome sequence indicates that plants and animals have not evolved by elaborating the same general process since separation from the last common ancestor. For example, although plants and animals have

comparable processes of pattern formation and the underlying genes appear to be similar, the actual mechanisms of getting to the end points of development are different. Once again, this reinforces the need to look specifically at the plant processes in order to understand how plants function.

One of the important ways in which the whole genome approach has changed plant biology is that international cooperation in many of the major projects is both necessary and important. The funding required for large-scale genomic sequencing makes it more important than ever to avoid unnecessary duplication. Thus the international coordination of both the *Arabidopsis* and the rice genome projects has ensured their completion with the minimal overlap of expenditure from the various international members, while still generating the appropriate scientific infrastructure and, in some cases, being responsible for the development of additional human and technological resources. These collaborations, both international as well as national, have improved the infrastructure for the science as well as moving knowledge forward at an ever-increasing rate.

The other important aspect of these genomics investigations is that the results are generally being widely disseminated, especially through Internet resources. Therefore, the constituency that is able to use these results to build detailed knowledge in specialist areas is ever widening. The structure of the informatics resources and the tools to query them must be compatible with the wide range of expertise of the interested parties. For individual investigators to be able to access and interrogate the results of major resource generators, such as sequencing projects, mutant collections, and the like, the data and resources must be made available. The availability of these resources is not just limited to the time that they are being actively generated but also after these projects are completed. Therefore, the archiving of biological and informatics resources to ensure their continued availability is vital, considering the investment that is being made in their generation.

The application of all this knowledge to the improvement of crops is not without controversy. The ability to manipulate plants for specific purposes with the introduction of new genetic material, that may or may not be of plant origin, is viewed with varying degrees of concern across the world. It is undoubtedly true that all of this new information can be useful in the development of new varieties by traditional breeding, but it will also have an input in developing totally novel strategies, including the use of plants to produce new raw materials. It will be important that the benefits of such engineered resources are spread across society and throughout the world to benefit both developed and developing countries, or they will never be generally accepted.

The primary aim of this text is to introduce the reader to the range of molecular techniques that can be applied to the investigation of unique and interesting facets of plant growth, development, and responses to the environment. The rapid progress made in this area has clearly been as a result

of increased funding in both the private and public sectors. The public sector efforts in the USA have been stimulated and supported by the National Plant Genome Initiative formally organized in 1997, along with major investments worldwide. This kind of support will be necessary for years to come to manipulate crop plants for improved productivity and ensure food security. The end result of all this investment should be a quicker introduction of new crop varieties in response to particular needs. The understanding of disease resistance, for example, and the development of new approaches to this problem are expected to reduce the time for new resistant varieties to be developed compared with the conventional introgression of new resistance genes from wild relatives. The combination of resources and technology that are currently available makes this an incredibly exciting time to be involved in plant genomics.

CHAPTER

1

THE STRUCTURE OF PLANT GENOMES

There is probably no one example that can be considered as the typical plant genome. They come in an amazing variety of shapes and sizes if one considers that the packaging into chromosomes is a form of shape. This variety can exist even within a family, with the result that plants are much more variable than any other group of organisms as far as these nuclear characteristics are concerned. In this chapter we consider how variable the DNA quantity can be, the variety of chromosome structures, and how all this variability in DNA quantity and packaging arose. These factors impinge on the design, feasibility, and interpretation of genomics studies.

DNA VARIATION—QUANTITY

The characteristic nuclear DNA value in a plant is generally expressed as the amount contained in the nucleus of a gamete (the 1C value), irrespective of whether the plant is a normal diploid or a polyploid (either recent or ancient). The use of a standard tissue is important because the nuclear DNA content can vary among tissues with some, for example the cotyledons of peas, having cells that have undergone many rounds of endoreduplication (Cullis and Davies, 1975). Nuclear DNA values have been reported in two different ways, either as a mass of DNA in picograms per 1C nucleus or as the number of megabase pairs of DNA per 1C nucleus. The relationship between these two ways is relatively easy to estimate because 1 pg of DNA is approximately equal to 1000 Mbp (the actual conversion is 1 pg ≡ 980 Mbp).

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

This 1C value for the amount of DNA in a plant nucleus can vary enormously. For example, one of the smallest genomes belongs to *Arabidopsis thaliana*, with 125 Mbp, whereas the largest reported to date belongs to *Fritillaria assyriaca*, with 124,852 Mbp, equivalent to 127.4 pg. This represents a 1000-fold difference in size between the largest and smallest genomes characterized so far. Some representatives that span these extremes are included in Table 1.1 and are taken from the database maintained by the Royal Botanic Gardens, Kew (http://www.rbgkew.org.uk/cval/homepage.html).

However, this range may not represent the true limits because DNA values have been estimated in representatives of only about 32% of angiosperm families (but only representing about 1% of angiosperm species), 16% of gymnosperm species, and less than 1% of pteridophytes and bryophytes. This variation occurs not only between genera but also within a genus. One example is the genus *Rosa*, in which there is a more than 11-fold variation in genome size. The fact that this range in DNA content is not associated with variation in the basic number of genes required for growth and development has led to its being referred to as the C-value paradox.

Genome size is an important biodiversity character that can also have practical implications. One example is that the genome size seems to constrain life cycle possibilities, in that all of those plants that have above a certain DNA content are obligate perennials (Bennett, 1972). Another example is that species with large amounts of DNA (>20 pg per 1C) can be problematic when studying genetic diversity with standard amplified fragment length polymorphism (AFLP) techniques such as have been encountered with *Cypripedium calceolus* (1C = 32.4 pg) and *Pinus pinaster*

TABLE 1.1. Selected DNA Values

Genus	Species	1C pg
Cardamine	*amara*	0.06
Arabidopsis	*thaliana*	0.125
Rosa	*wichuraiana*	0.13
Luzula	*pilosa*	0.28
Oryza	*sativa*	0.5
Rosa	*moyesii*	1.45
Gnetum	*ula*	2.25
Zea	*mays*	2.73
Nicotiana	*tobaccum*	5.85
Ginkgo	*biloba*	9.95
Allium	*sativum*	16.23
Pinus	*ponderosa*	24.2
Fritillaria	*assyriaca*	127.4

From http://www.rbgkew.org.uk/cval/homepage.html

(1C = 24 pg) (cited in Bennett et al., 2000). On the other hand, a very small DNA content has been a major factor in determining the early candidates for genome sequencing. Consequently, *Arabidopsis thaliana* (a dicot) was the first plant chosen for genome sequencing, partly because it had one of the smallest C values known for angiosperms. Rice was the second genome sequenced and was the first monocot chosen because it had the smallest C value among the world's major cereal crops, even though it did not have the smallest genome in the grasses. This distinction currently goes to the diploid *Brachypodium distachyon*, which has a 1C value of 0.25–0.3 pg, whereas the rice genome is nearly twice this size (Bennett et al., 2000).

The determination of the genome sequence of *Arabidopsis* gives some indication of what the minimum genome size for a higher plant is likely to be. The extensive duplication that was found in the *A. thaliana* genome could well have been the result of polyploidy earlier in the evolutionary history of this plant. Thus the number of genes necessary and sufficient to determine a functional higher plant is likely to be somewhat less than 25,000, the current estimate for *A. thaliana*. Additional DNA will need to be associated with these genes to ensure appropriate chromosome function by defining the centromeres and telomeres. Therefore, the most stripped-down plant genome is unlikely to be much below 0.1 Gb, because in addition to the 25,000 genes, DNA associated with centromeres and telomeres that ensure chromosome stability and segregation at cell division will also have to be included. However, a great deal more information is still required before a conclusion that this minimal number will be sufficient to ensure the full range of functions that can be performed by plants.

As will be seen below the actual amount of DNA that is associated with various structures within the genome can vary. However, it is not just in this context that it is important to know the C value. DNA amounts have been shown to correlate with various plant life histories, the geographic distribution of crop plants, plant phenology, biomass, and sensitivity of growth to environmental variables such as temperature and frost. The C value may also be a predictor of the responses of vegetation to man-made catastrophes such as nuclear incidents. It has been shown that plants with a higher DNA content and particular chromosome structures are more resistant to radiation damage (Grime, 1986).

Chromosome Variation

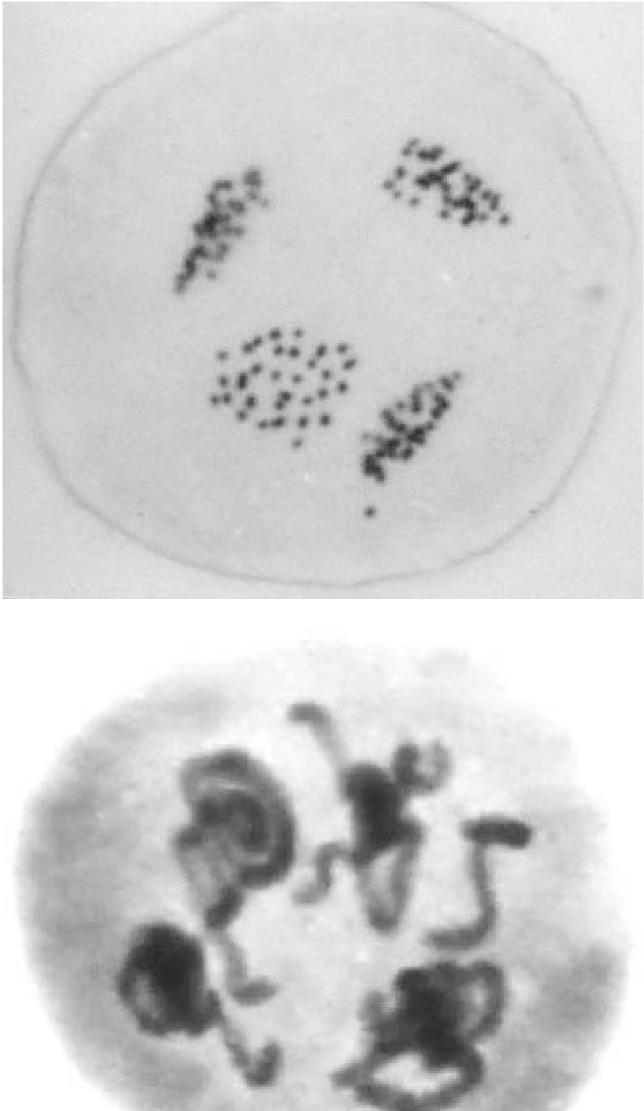
Chromosome number and size are very variable. The stonecrop, *Sedum suaveolens*, has the highest chromosome number (2n of about 640), whereas the lowest chromosome number is that of *Haplopappus gracilis* (2n = 4). Ferns also have extremely high values. An increase in the number of chromosomes is usually associated with a reduction in chromosome size. The actual

structure of a chromosome can also vary, with most species having the usual chromosome structure of a single centromere. However, some plants have holocentric chromosomes where kinetochore activity (regions that attach to the spindle at mitosis and meiosis) is present at a number of places all along the chromosome.

In the genus *Luzula*, which has holocentric chromosomes, the chromosome number can also vary widely, with *L. pilosa* having 66 chromosomes and *L. elegans* having 6 as the diploid number (Figure 1.1a, b). As can be seen in the figure, the size of a chromosome in these two species is very different. The quantity of DNA in each chromosome is also very different; *L. elegans* has 3 chromosomes in which to package the 1446 Mbp of DNA in the 1C nucleus, whereas in *L. pilosa* 33 chromosomes are available for only 270 Mbp of DNA in the 1C nucleus. Each of the *L. elegans* chromosomes is of similar size and contains an average of 482 Mbp of DNA, whereas each *L. pilosa* chromosome only packages about 8 Mbp of DNA. Therefore, within this genus, a single chromosome of one species (*L. elegans*) contains an amount of DNA equivalent to that present in the complete rice genome, whereas the other (*L. pilosa*) has chromosomes that are each the size of an average microbial genome.

The arrangement of kinetochore activity all along the chromosome has consequences for meiosis, including a restriction of the reduction division to the second division of meiosis rather than the first, as is the case in most plants. It also restricts the regions that can recombine and so may have other consequences for the plants that must be considered in relation to function and evolution of the genome. However, it does mean that almost any chromosome fragment will have a kinetochore and so be maintained through cell division. Therefore, fragmentation of the chromosomes will not be lethal and can generate different chromosome numbers. The organization of the genome into this type of package leads to extreme resistance to radiation damage. Figure 1.2 shows mitosis from a callus cell of *L. elegans*. Although the plants were grown from irradiated seeds they showed no apparent phenotypic abnormalities. In fact, plants are very tolerant of chromosome aberrations, with ploidy changes being very frequent. This property can be utilized in generating material that is targeted to understanding of particular regions of the genome, for example, the production of wheat addition and deletion lines that have been important resources in the effort to unravel the enormous wheat genome (Sears, 1954) and for isolating single maize chromosomes (Kynast et al., 2001).

As mentioned above for the genus *Luzula*, the chromosomes can vary greatly both in size and number. Situations also exist in which there is relatively little difference in the chromosome number but there are very large differences in the chromosome sizes. Within the legumes this has been extensively characterized. For example, both *Vicia faba* and *Lotus tenuis* have a chromosome number of 6, whereas the lengths of these



a

b

FIGURE 1.1. Metaphase of meiosis II in *L. pilosa* (a) and *L. elegans* (b). (Photographs by Dr. G. Creissen.)

FIGURE 1.2. Mitotic metaphase in *L. elegans* callus derived from seed irradiated with 80 krad. At least 3 centric fragments are visible in addition to the 6 chromosomes. (Photograph by Dr. B. Bowen.)

TABLE 1.2. Chromosome Number, Chromosome Length, and DNA Content of Two Legumes

Species	Haploid set of chromosomes (n)	Average length of chromosomes (μm)	Nuclear DNA content (pg)
Lotus tenuis	6	1.8	0.48
Vicia faba	6	14.8	13.33

From http://www.biologie.uni-hamburg.de/b-online/e37/37c.htm

chromosomes only partly reflect the differing DNA contents in these two species (Table 1.2, Figure 1.3), with the DNA per unit length differing over threefold (0.044 pg/μm in *Lotus* and 0.15 pg/μm in *Vicia*) (from http://www.biologie.uni-hamburg.de/b-online/e37/37c.htm).

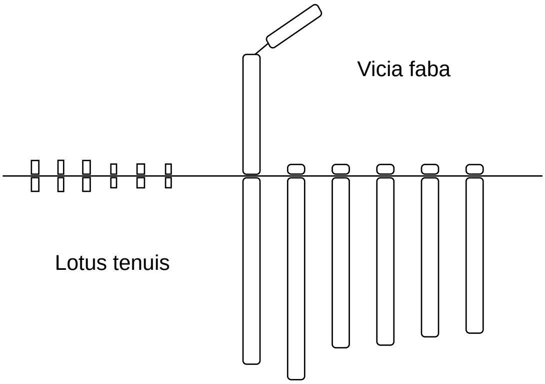


FIGURE 1.3. Chromosome sizes in *Lotus* and *Vicia*.
(From http://www.biologie.uni-hamburg.de/b-online/e37/37c.htm.)

Origin of DNA Variation

The sequences in the genome are generally classified with respect to the number of times they are represented. The three main classes to which they are assigned, low copy, moderately repetitive, or highly repetitive, have somewhat arbitrary cutoffs, with both copy number and function playing a part in the classification. These three classes and some of their characteristics are:

- Low-copy-number or unique sequences that probably represent the genes
- Moderately repetitive sequences, many of which may be members of transposable element families that are distributed around the genome
- Highly repetitive sequences, many of which are arranged in tandem arrays

The arrangement of these sequences with respect to one another has functional consequences for the plant.

Low-Copy Sequences

The two complete genome sequences from *Arabidopsis thaliana* and rice are from genomes that vary nearly fourfold in size, so the estimates of gene number from these two sequences will go some way toward establishing how the gene number might change with genome size. The initial estimates from the rice genome sequence (Goff et al., 2002) are that rice has about twice the number of genes that are found in *Arabidopsis*. As gene finding programs

continue to improve, this number in rice may well decrease, and so the most likely trend is that approximately the same number of genes will be present in all plants irrespective of the total amount of DNA in the nucleus. The question of how a gene is defined will keep cropping up. Are all the members of a gene family counted as a single gene, or is each member an individual gene? How different do the members of a family have to be to be counted as different genes? How similar do the sequences, or the protein domains, need to be for the genes to be placed in a family? One extreme example is the family of genes encoding the protein ubiquitin. This protein is probably the most conserved protein, at the amino acid level, across virtually all eukaryotes, but adjacent members in a flax polyubiquitin differed by 24% in their nucleic acid sequence although the amino acid sequence of the members was identical (Agarwal and Cullis, 1991).

Arabidopsis has many more gene families with more than two members than has been found in other eukaryotes (The Arabidopsis Genome Initiative, 2000). These families are generated in a number of different ways. Segmental duplication, that is, the presence of a segment of one chromosome somewhere else in the genome with a series of genes present within the segment, is responsible for more than 6000 gene duplications. Higher copy numbers (that is >2, the number generated by the segmental duplications) of genes within a family are frequently generated by tandem amplifications, where the gene is either repeated many times within a stretch of the genome or spread through the chromosome complement. An example of this amplification is seen in the genes for the storage protein zein in maize, where a 78-kbp region of the maize genome contains 10 related copies of a 22-kDa zein gene (Song et al., 2001). The complete genome sequences of *Arabidopsis* and rice show many local tandem amplifications. For example, an analysis of the BAC clone F16P2 from *Arabidopsis* has three gene families, glutathione-*S*-transferase and tropinone reductase genes and a pumilio-like protein present as tandem arrays as shown in Figure 1.4 (Lin et al., 1999). In rice the GST gene has 63 recognizable copies, 23 of which are located on chromosome 10L. Sixteen additional GST genes are present in three other clusters located near the centromere of chromosome 1 (8 genes) and on 1L (4 genes) and 3S (4 genes) (Yuan et al., 2002).

Analysis of the *Arabidopsis* genome sequence has revealed arrays of various individual genes ranging up to 23 adjacent members and containing 4140 individual genes. This represents 17% of all genes of *Arabidopsis* that are arranged in tandem arrays. The high proportion of tandem duplications also indicates that unequal crossing over is the likely mechanism by which new gene copies are generated (The *Arabidopsis* Genome Initiative, 2000). This feature of the *Arabidopsis* genome, which would also be expected to be present in other plant genomes, is consistent with a relaxed constraint on the genome size in plants allowing tandem duplications without disruption of the control of gene expression.

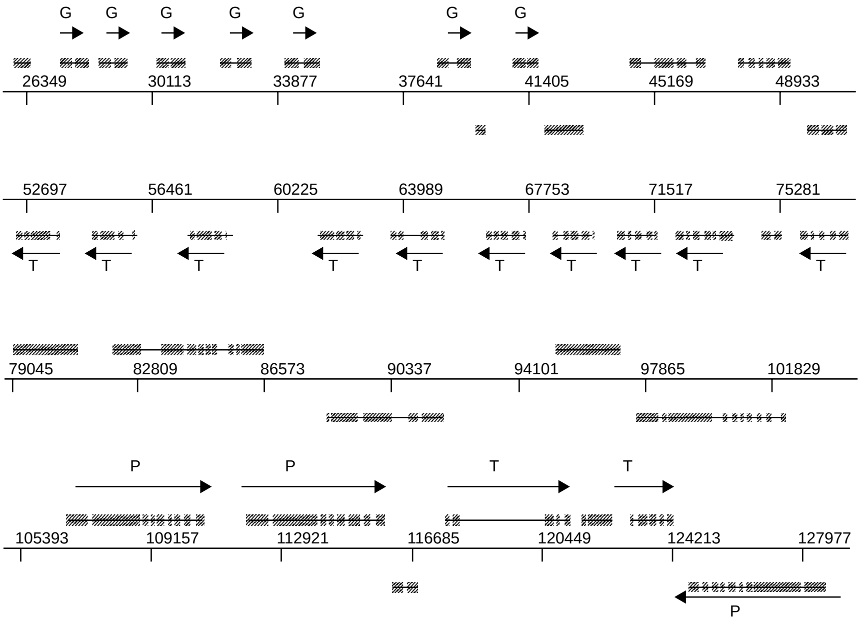


FIGURE 1.4. Organization of genes on BAC F16P2 showing the 3 tandem gene duplications. The display from TIGR Annotator shows the exon/intron structure of the annotated genes. The glutathione-*S*-transferase and tropinone reductase genes are labeled G and T, respectively. A smaller duplication of pumilio-like protein (P) is also present (This image is provided courtesy of The Institute for Genomic Research (TIGR), 9712 Medical Center Dr., Rockville, MD 208850. The original published figure and the scientific details of the research can be found in Nature 1999 December 16; 402:761–767).

The high degree of duplications, but not triplication, of large chromosomal segments makes it most likely that *Arabidopsis*, like many other plant species, had a tetraploid ancestor with subsequent divergence, loss, and reassortment of the tetraploid genome. However, it is also possible that the duplicated segments were the result of many independent duplication events rather than being the result of tetraploid formation.

A question arises concerning how one counts the gene number. Are duplicated sequences counted as a single gene even if the sequence has diverged but still contains an open reading frame? As the genome increases in size many gene-containing regions will also be duplicated or arise at higher multiplicities. If these genes diverge and as a consequence gain a new specificity, should this be counted as an additional gene? If so, then it is possible that the number of genes will rise as the genome gets bigger. For example, in *Arabidopsis* genomic analysis of the terpenoid synthase

gene family has revealed a set of 40 genes that cluster into five superfamilies (Aubourg et al., 2002). Are these to be counted as a single gene, five genes, forty genes, or thirty-two genes, as eight are interrupted and likely to be pseudogenes? Even one of these putative pseudogenes is present in the collection of EST sequences so that even transcription may not be a sufficient discriminator.

The evidence from the complete genome sequences of *Arabidopsis* and rice make it abundantly clear that all the extra DNA in rice does not represent genes. In general, the extra DNA is made up of repetitive sequences. These repetitive sequences can be of two types, either dispersed through the genome or present in tandem arrays of a unit repeat.

Dispersed Repetitive Sequences

The dispersed repetitive sequences are generally thought to be derived from transposable elements. As the genome size increases, so does the proportion of the genome that is recognizable as being related to these transposons. Transposons have been found in all eukaryotes and prokaryotes and can be of two types:

- Class I—These are retrotransposons that replicate through an RNA intermediate and so increase in number with each round of transposition.
- Class II—These are transposons that move directly through a DNA form and so move position without normally increasing in number.

Evidence has been accumulating that the genome size variation is correlated with both the number of different retrotransposon families and the level of retrotransposons present in the genome. This situation seems to be especially true in the grasses (Bennetzen, 1996).

About 10% of the *Arabidopsis* nuclear DNA is present in the form of transposons even though *Arabidopsis* has a relatively compact and simple genome (The *Arabidopsis* Genome Initiative, 2000). On the other hand, maize has literally thousands of different families of retrotransposons. These retrotransposons themselves can be divided into two categories, those that contain long terminal repeats (LTR) at the ends of the transposon and those that do not. The retrotransposons that have a similar structure and conserved LTR sequences are thought to belong to families derived from a common element. The retrotransposons are frequently present in clusters in the intergenic regions. An example of such clustering of transposon sequences is an intergenic region in maize that was found to have nested retrotransposons representing 10 different families (Figure 1.5). Each of these families was also present elsewhere in the genome, with a total of 10,000 to 30,000 copies. These repeats, that is, transposons, represented 60% of the total DNA within

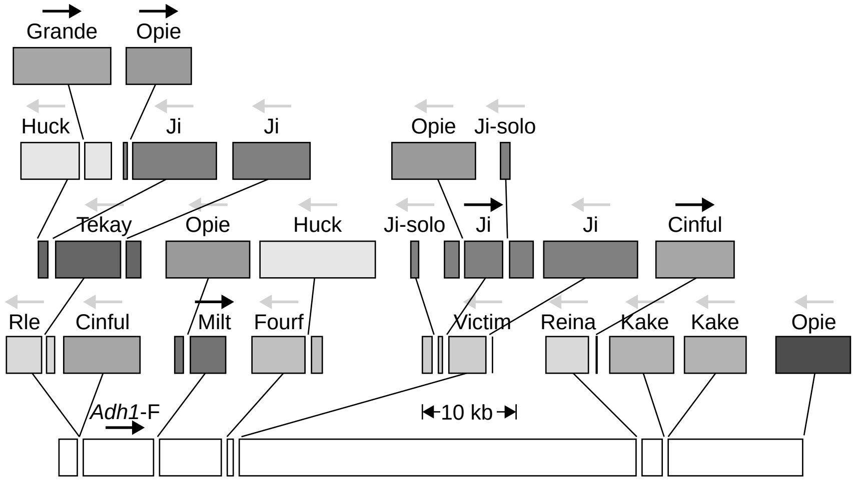


FIGURE 1.5. The structure of the *Adh1*-F region of maize, showing identified retrotransposons. Only one gene is shown (*Adh1*-F), although more genes are present on this segment. The arrow above each element indicates its orientation. (Figure provided by Dr. J. Bennetzen.)

the sequenced 280 kbp spanning the original clone. Similar clusters of retroelements are dispersed throughout the maize genome (SanMiguel et al., 1996). This type of organization is expected to be seen throughout the grasses, especially those with larger genomes. However, within the rice genome (one of the smaller genome grasses) miniature inverted repeat transposable elements (MITES) seem to be more prevalent and the number of families and copy number of elements in each family are much lower (Bennetzen, 2002). Is this because those genomes of smaller size prevent transposon explosions, thereby preventing the number from ever rising, or do they have more efficient expulsion/eradication/elimination mechanisms that effectively remove the newly amplified, or even established, copies?

Tandemly Repeated Sequences

The tandemly repeated sequences fall into at least three classes. These include centromeric satellite repeats that are located between each chromosome arm and span the centromere, the telomeric regions, and the ribosomal RNA genes. The ribosomal RNA genes coding for the large ribosomal RNAs are the longest tandem repeated sequences, with a repeat length of about 10 kb. Most of the remaining families tend to be about either 180 or 360 bp long. These lengths are similar to multiples of the unit length of DNA in a nucleosome, and the unit length itself may be more important than the actual nucleotide sequence.

Centromeric DNA mediates chromosome attachment to the meiotic and mitotic spindles and often forms dense heterochromatin. The *Arabidopsis* genome sequence has identified the centromeric regions, which contain numerous repetitive elements including retroelements, transposons, microsatellites, and middle repetitive DNA. An unexpected observation was that at least 47 expressed genes were encoded in the genetically defined centromeres of *Arabidopsis* (Copenhaver et al., 1999). The regions containing these repeats also contain many more class I than class II elements (Figure 1.6). Because few centromeres, in fact only those from *Arabidopsis* and rice,

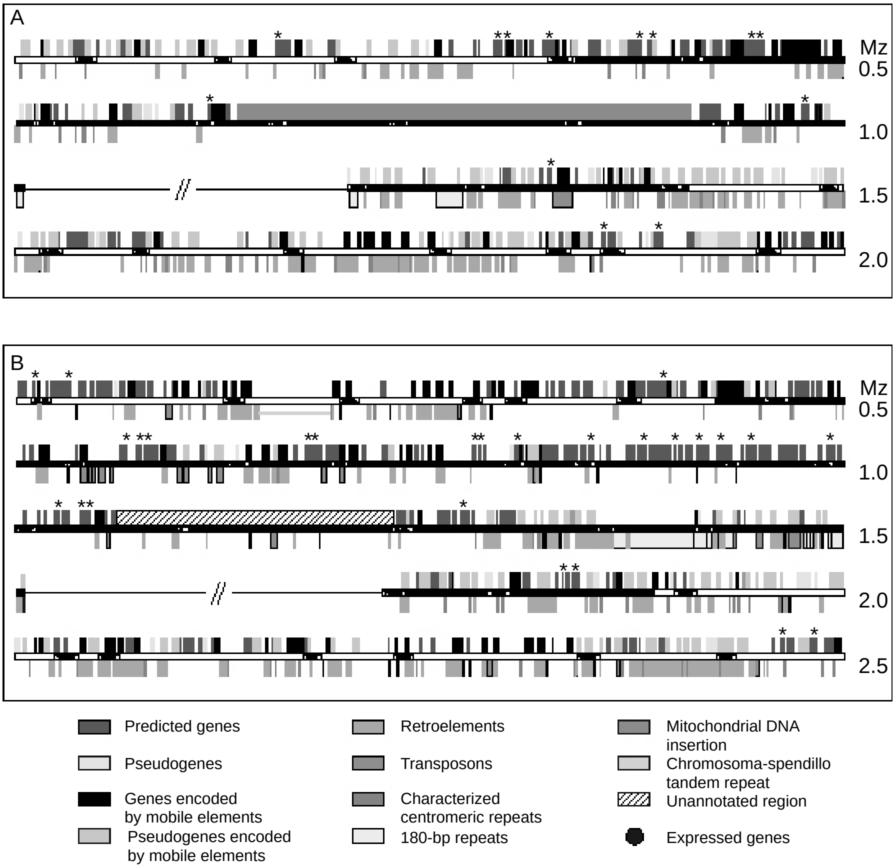


FIGURE 1.6. Sequence features at *CEN2* (A) and *CEN4* (B). Central bars depict annotated genomic sequence of indicated BAC clones; black, genetically defined centromeres; white, regions flanking the centromeres; //, gaps in physical maps. Sequences corresponding to genes and repetitive features, filled boxes (above and below the bars, respectively). (Reprinted with permission from Copenhaver et al., Science **286**, 2468–2474. Copyright (1999) American Association for the Advancement of Science.)

have been identified, the general structure of a centromere still must be determined. Another unanswered question relates to the structure of the kinetochore in comparison to the centromere. Will a kinetochore have an attraction for transposons similar to that seen for the *Arabidopsis* centromere, and so have a complex structure, or be a simpler stripped-down attachment site, like that of yeast, that will make it easier to understand the essential functions necessary for chromosome movement? There is evidence for conserved and variable domains among the centromere satellites from *Arabidopsis* populations (Hall et al., 2003).

The genes encoding the 18S, 5.8S, and 25S ribosomal RNAs are present in tandem arrays of unit repeats in a recognizable chromosome structure, the nucleolar organizer region (NOR). The repeat unit consists of the coding sequences for each of these three RNAs as well as an internal transcribed spacer region and an intergenic region (Figure 1.7). The number of repeating units varies between several hundred and over 20,000. Therefore, a plant that has 20,000 copies of the ribosomal RNA genes has almost as much DNA in this one tandemly arrayed family as *Arabidopsis* has in its whole genome. The number of repeat units of these genes varies within a species and may even vary within a plant (Rogers and Bendich, 1987). Even between maize inbred lines the variation is more than twofold (Rivin et al., 1986). The variation in this gene family would account for a DNA difference of about 100 Mbp. Gymnosperms have a much longer repeat unit than angiosperms (Cullis et al., 1988).

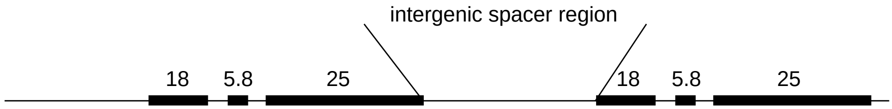


FIGURE 1.7. The repeat unit for the large ribosomal RNA genes.

PROCESSES THAT AFFECT GENOME SIZE

The genome can be extensively amplified by duplicating either part or all of the genome through polyploidy. Polyploids have more than two complete sets of chromosomes in their nuclei compared with the two that are found in normal diploids. The rate of polyploidy in different groups is variable and has been estimated as up to 80% in angiosperms, 95% in pteridophytes, but relatively uncommon in gymnosperms. Polyploidy can arise in two different ways (Figure 1.8). One of these is by doubling the chromosomes of a single individual resulting in autopolyploidy. The other is by combining the genomes from two closely related species. This latter event, which frequently happens in a wide cross, results in the genomes of two different species residing in the same nucleus (allopolyploid). If the chromosomes from the two genomes have diverged sufficiently so that the homologs from the two

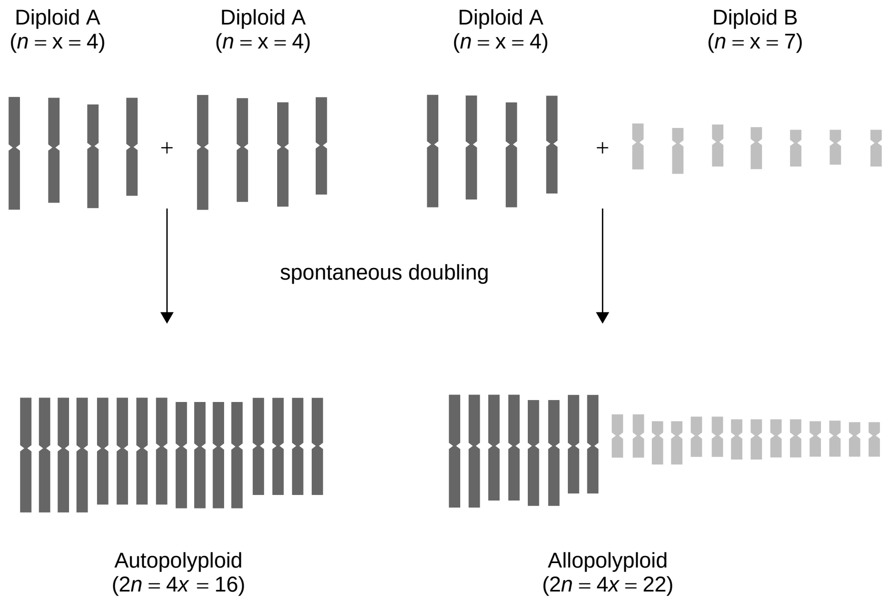


FIGURE 1.8. Mechanisms of polyploidization. (Reprinted from Trends in Pl. Sci. 2, Leitch and Bennett, Polyploidy in Angiosperms, 470–476, Copyright (1997), with permission from Elsevier.)

species do not pair efficiently at meiosis, then the hybrid will be sterile. However, a doubling of the chromosome number will result in a normal meiosis and a new polyploid species will have been formed. Polyploids are very frequent in the angiosperms, and most of the major crop species are polyploids. These rounds of polyploidization are insufficient to account for all of the increase in genome size seen in the angiosperms. To see an increase of a thousandfold in the DNA content would require approximately 10 sequential rounds of doublings to have occurred. Octaploids seem to be the largest frequently observed polyploids, only representing three sequential doublings. However, the stonecrop is estimated to be about 80-ploid with about 640 chromosomes (Leitch and Bennet, 1997). Despite this upper value, the largest genomes are not the result of many rounds of whole genome doublings.

Rather than the addition of a complete genome, various mechanisms can result in the duplication of large regions of the genome. These mechanisms include unequal recombination and nonreciprocal translocations. Both of these mechanisms would result in one product having a loss of DNA while the other has an increase. There would have to be a selective advantage for the product that had a duplication in order for the genome to grow by this method. Again, as pointed out for polyploidy, the number of rounds of

duplications needed to grow the genomes to the size that are seen are much greater than would be supported by our current estimations of gene family sizes. Therefore, other processes need to be operational apart from whole and partial genome duplications.

Most of the genome size increases in the grasses appear to be the result of the amplification of retrotransposable element families. The retrotransposons can increase their numbers within the genome because transposition acts through an RNA intermediate. This therefore leaves the original DNA copy in the genome, while placing additional copies elsewhere in the genome. These elements should be acted on by natural selection so that their continued expansion has led to their being named selfish or parasitic elements. The dispersal of rogue RNA polymerase III transcription products, such as the Alu elements in humans and perhaps the expansion of the 5S ribosomal RNA genes in flax, are demonstrations of this behavior. In all the plants in which they have been investigated the LTR retrotransposons are the biggest variable that can be related to genome size. These elements can make up 60% or more of genomes like maize, wheat, and barley but less than 50% in rice and around 10% in *Arabidopsis*. The rice genome also contains numerous inverted repeat transposable elements such as the MITES, which, although numerous, are too short to have a large impact on the overall genome size.

The rounds of polyploidy and segmental duplications, along with transposable element family amplification, all result in an increase in genome size (Table 1.3). So are plant genomes destined to hold a one-way ticket to

TABLE 1.3. A Hypothetical Case of the Effects of Genome Expansion and Contraction

Event	Diploid chromosome number	2C DNA content
Ancestral genome	4	1 pg
Autopolyploidy	8	2 pg
Divergence, loss of DNA sequences	8	1.9 pg
Retrotransposon explosion	8	2.3 pg
Alloploidy with species containing 6 chromosomes as diploid number and 2 pg of DNA/2C nucleus	20	4.3 pg
Retrotransposon explosion	20	5.3 pg
Loss of chromosome pair (one of pairs containing ribosomal RNA genes)	18	5.1 pg

"genomic obesity" (Bennetzen and Kellogg, 1997), or are there ways the genome can be decreased? The loss of genes from polyploids has been observed, sometimes at very high rates. These losses are associated with deletions that are much smaller than the loss of whole chromosomes (Levy and Feldman, 2002).

Because much of the variation in genome size is associated with retrotransposons, their removal could be an important factor in downsizing the genome. Unequal recombination mechanisms can remove retrotransposon sequences because the LTR regions are in a direct orientation and share a very high degree of sequence homology. Therefore, if within a region of the chromosome there were a number of insertions of related retrotransposon sequences, then recombination between the two ends of this array would result in a deletion of the array with the generation of a single LTR with no other detectable associated LTR to define an intact element. The BARE-1 retroelement in barley demonstrates this phenomenon. The relative ratio of solo LTRs to intact elements in barley and its wild relatives is consistent with this model for retroelement copy number reduction (Bennetzen, 2002).

Consequences of Multiple Genomes

How do plants cope with all of this extra DNA in the genome? Of particular interest are the mechanisms by which the genes still function appropriately in a newly formed polyploid. The phenomena of gene silencing has been clearly demonstrated when an additional gene copy has been added by transformation as well as naturally occurring examples in polyploid wheat. In polyploids gene silencing was first observed for the ribosomal RNA genes. The epigenetic phenomenon called nucleolar dominance that results in the complete silencing of one parental set of rRNA genes in a genetic hybrid, or the silencing of the particular nucleolar organizer regions in the grasses, is an extreme example of gene silencing (Flavell et al., 1993). Even within a cluster of ribosomal RNA gene repeats, all of the copies may not be expressed, so this example may have an expression control mechanism that may be utilized for silencing it or may be totally independent of any silencing mechanism(s).

The copy number of these genes varies greatly between species or genera, but a much narrower range of values is found within a species. The ribosomal RNA gene copy number in the genome is thought to be modulated by unequal crossing over. It is not known whether the molecular processes involved in the control of expression have any effect on this copy number variability.

Syntenic relationships are the relative placement of genes with respect to one another in different species. Therefore, synteny is a measure of the chromosomal rearrangements since the divergence from a common ancestor

when the chromosomal distribution of genes in different species is compared. The development of molecular maps that have included many genes has allowed the spacing and/or ordering of these genes in different species to be compared. The identification of shared chromosome regions, or even whole chromosomes, in terms of the genes present and their order, in many plants has been determined (Figure 1.9). The level of synteny allows an estimate of the number of rearrangements required to account for the patterns seen. In many comparisons, large segments of chromosomes (or sometimes entire chromosomes) are found to have the same order of genes. However, the spacing between mapped genes, even in molecular maps, is not always proportional.

These syntenic relationships have been useful in understanding the evolutionary processes that may have occurred in plants as well as in

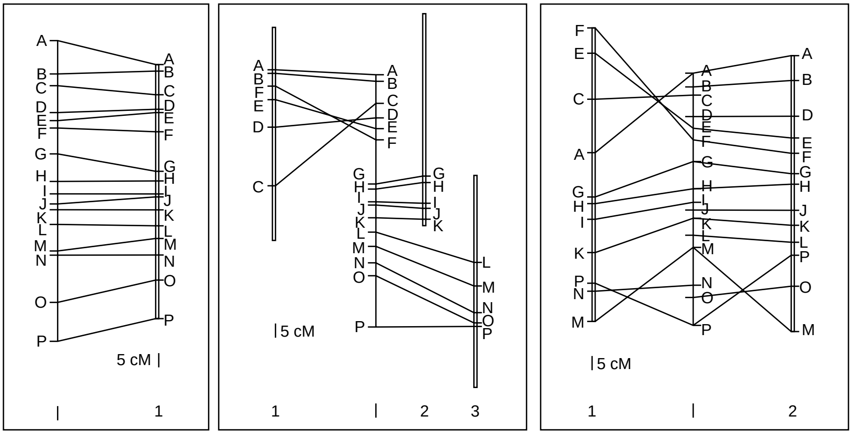


FIGURE 1.9. Patterns of genome collinearity. The use of the same set of molecular markers (A–P) for genetic mapping experiments in different species allows the alignment of the resulting chromosome maps. In the left part of the figure, 2 chromosome maps (| and 1) are shown, which are completely collinear. The central part of the figure outlines the case in which a chromosome from a particular species (|) shares collinear segments with several chromosomes of another species (1–3) indicating translocation events. Inversions of entire chromosome arms or smaller chromosomal segments are also frequently observed in comparative genetic mapping experiments. If a diploid and a tetraploid species are compared, markers will generally reveal 2 loci in the tetraploid species. In the right part of the figure, chromosomes 1 and 2 of a tetraploid species are aligned with chromosome | of the diploid species. Depending on the degree of polymorphism between the 2 species analyzed, not all of the markers will reveal 2 different loci in the tetraploid species, as indicated for example for markers B and N. (Reprinted from Current Opinion in Plant Biology 3, Schmidt, Synteny: Recent Advances and Future prospects. 97–102, Copyright (2000), with permission from Elsevier.)

other phyla. Thus the large-scale rearrangements of chromosome segments have occurred rarely during evolution so that the deciphering of the order of rearrangements should assist in understanding evolutionary relationships. However, it has been observed that rearrangements have occurred more frequently in some evolutionary lineages than in others. Syntenic relationships are the relative placement of genes with respect to one another. For these relationships to be observed order, rather than physical distances, must have been conserved in plants that vary greatly in genome size. Therefore, the genome contraction or expansion events must have occurred more frequently than rearrangements for size to vary but order to remain relatively constant.

Aligned maps might be exploited to identify many different markers from a variety of species for a given genomic region. This could be especially useful for fine-scale mapping or map-based cloning experiments. Knowing a little about the linkage of a desirable trait in an economically important but not well-studied organism would allow the examination of syntenic segments of a better-studied organism to identify genes that are candidates for the trait. However, the level of microsynteny (the exact linear arrangement of genes within the chromosomal segment) does not appear to be as faithful as that of macrosynteny (the presence of a cluster of genes within a chromosomal region). The detailed order of genes within a syntenic region may be much more variable than the clustering of genes within a region of the chromosome. Therefore, this approach for candidate gene cloning may be fraught with peril.

Obviously, the occurrence of polyploidy will affect any syntenic relationships that may be discovered. Because there ought to be very closely related or identical genomes within the polyploid nucleus the duplicated segments should be virtually identical. As the two genomes diverge after the initial genome amplification event, it will become more and more difficult to identify which particular segment is syntenic to one from a different species. Both ancestral regions will clearly share homology, and additional information will be required to identify a functionally equivalent region. Also, because two copies of each gene will be present, the one that is functionally equivalent will also be more difficult to identify. Therefore, the notion of paralogs and orthologs has been introduced. The orthologs are copies of the genes that are functionally equivalent, whereas the paralogs are related in sequence but not necessarily of identical function.

Maize, sorghum, and sugarcane have been intensively studied for conservation of linkage arrangements (Ramakrishna et al., 2002). The sorghum and sugarcane genomes showed very similar linear arrangements of related features along the chromosome. However, when these two genome arrangements were compared with that of maize, two different regions in the maize genome frequently showed homology to a single region in the sorghum and sugarcane genomes. The observation of these duplicated regions is consis-

tent with the view that maize is an ancient tetraploid compared with sorghum and sugarcane, which are still diploid. The level of microsynteny for the genes themselves may vary depending on the region of the chromosome being investigated. For example, a lack of synteny in the regions containing pathogen resistance-like genes may be observed even in closely related plants because these regions appear to be rapidly evolving, whereas for other regions in the same comparison the result may be high level of microsynteny.

What is the end result of all these changes in genome size on the organization of genes in large and small genomes? Because the consensus of opinion is that the number of genes is approximately the same in all plants, the gene density along the chromosome must be much lower in large genomes compared to small genomes. Also, as much of the increased DNA is in the form of transposable elements inserted between genes, the picture that emerges is that as the genome increases in size, the density of genes per unit chromosome length decreases, with many more repetitive elements being present between each of the genes. However, some regions of the genome appear to be "sinks" for transposable elements. For example, the centromeric region in *Arabidopsis* has a much higher density of transposons than other regions of the genome. If the same distribution of transposons occurs in large genomes, then the relative separation of individual genes in these genomes may not be as great as the increase in the DNA content would initially indicate. This leads to the concept of gene-rich regions, that is, regions that are much higher in gene content than expected, and also the necessary presence of gene-poor regions. There is evidence for such gene-rich regions, especially in large cereal genomes. The presence of such gene-rich regions will obviously affect sequencing strategies if the aim is to preferentially identify genes rather than large stretches of transposons, repeats, and other nontranscribed sequences. These gene-rich regions will not contain all the genes, as has been demonstrated in *Arabidopsis*, where there are genes within the centromeric region. However, because the proportion of genes in this region of the genome is very small, the strategy of targeting gene-rich regions may not miss many of the interesting genes.

SUMMARY

The genome itself has a level of biological organization with its own phenotype forged by a complex evolutionary history. More importantly, there is an intimate relationship between the phenotype of the genome and the phenotype of the organism. Variation in genome size can constrain, and simultaneously be constrained by, the evolution of the organism in the environment in which it finds itself. The remarkable degree of genome plasticity may be necessary to provide new functions for plants. The

emerging picture of the molecular organization of complex plant genomes is one of a mosaic with tracts of highly repetitive heterochromatin interspersed with regions of transposons and of transcriptionally active genes. A large proportion of the active genes may be clustered into gene-rich regions that are themselves separated by tracts of complex transposons. The context in which the genes exist, as well as the presence of multiple members of gene families, must be included in the consideration of the overall control of gene expression.

References

Agarwal, M. L. and C. A. Cullis (1991) The ubiquitin genes in flax. *Gene* **99**, 69–75.

Aubourg, S, A. Lecharny. and J. Bohlmann. (2002) Genomic analysis of the terpenoid synthase (Attps) gene family of *Arabidopsis thaliana*. *Mol. Genet. Genomics* **267**, 730–745.

Bennett, M. D. (1972) Nuclear DNA content and minimum generation time in herbaceous plants. *Proc. R. Soc. Lond. B* **181**, 109–135.

Bennett, M. D., P. Bhandol. and I. J. Leitch (2000) Nuclear DNA amounts in angiosperms and their modern uses—807 new estimates. *Ann. Botany* **86**, 859–909.

Bennetzen, J. L. (1996) The contributions of retroelements to plant genome organization, function and evolution. *Trends Microbiol.* **4**, 347–353.

Bennetzen, J.L. (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* **115**, 29–26.

Bennetzen, J. L., and E. A. Kellogg (1997) Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9**, 1509–1514.

Copenhaver G. P., K. Nickel, T. Kuromori, M. I. Benito, S. Kaul, X. Y. Lin, M. Bevan, G. Murphy, B. Harris, L. D. Parnell, W. R. McCombie, R. A. Martienssen, M. Marra, and D. Preuss (1999) Genetic definition and sequence analysis of *Arabidopsis* centromeres Science **286**, 2468–2474.

Cullis, C. A., G. P. Creissen, S. W. Gorman, and R.D. Teasdale (1988) The 25S, 18S, and 5S ribosomal RNA genes from *Pinus radiata*. In: *IUFRO Workshop on Molecular Biology of Forest Trees*. Ed. W. M. Cheliak and A. C. Yapa, Canadian Forestry Service, Petawawa, 34–40.

Cullis, C. A., and D. R. Davies (1975) Ribosomal DNA amounts in *Pisum sativum*. *Genetics* **81**, 485–492.

Flavell, R. B., M. Odell, R. Sardana, and S. Jackson (1993) Regulatory DNA of ribosomal-RNA genes and control of nucleolus organizer activity in wheat. *Crop Sci.* **33**, 889–894.

Goff, S. A., D. Ricke, T. H. Lan, G. Presting, R. L. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchinson, C. Martin, F. Katagiri, B. M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. P. Zhong, T. Miguel, U. Paszkowski, S. P. Zhang, M. Colbert, W. L. Sun, L. L. Chen, B. Cooper, S. Park, T. C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. S. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid,

S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R. M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant, and S. Briggs (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp japonica) *Science* **296**, 92–100.

Grime, J. P. (1986) Prediction of terrestrial vegetation responses to nuclear winter conditions. *Int. J. Environ. Studies* **28**, 11–19.

Hall, S. E., G. Kettler. and D. Preuss (2003) Centromere satellites from *Arabidopsis* populations**:** Maintenance of conserved and variable domains. *Genome Res.* **13**, 195–205.

Kynast, R. G., O. Riera-Lizarazu, M. I. Vales, R. J. Okagaki, S. B. Maquieira, G. Chen, E. V. Ananiev, W. E. Odland, C. D. Russell, A. O. Stec, S. M. Livingston, H. A. Zaia, H. W. Rines. and R. L. Phillips (2001) A complete set of maize individual chromosome additions to the oat genome. *Plant Physiol.* **125** 1216–1227.

Leitch, I. J., and M. D. Bennet (1997) Polyploidy in angiosperms. *Trends Plant Sci.* **2**, 470–476.

Levy, A. A., and M. Feldman (2002) The impact of polyploidy on grass genome evolution. *Plant Physiol.* **130**, 1587–1593.

Lin, X., S. Kaul, S. Rounsley, T. P. Shea, M-I. Benito, C. D. Town, C. Y. Fujii, T. Mason, C. L. Bowman, M. Barnstead, T. V. Feldblyum, C. R. Buell, K. A. Ketchum, J. Lee, C. M. Ronning, H. L. Koo, K. S. Moffat, L. A. Cronin, M. Shen, G. Pai, S. Van Aken, L. Umayam, L. J. Tallon, J. E. Gill, M. D. Adams, A. J. Carrera, T. H. Creasy, H. M. Goodman, C. R. Somerville, G. P. Copenhaver, D. Preuss, W. C. Nierman, O. White, J. A. Eisen, S. L. Salzberg, C. M. Fraser, and J. C. Venter (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**, 761–769.

Ramakrishna, W., J. Dubcovsky, Y. J. Park, C. Busso, J. Emberton, P. Sanmiguel, and J. L. Bennetzen (2002) Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics* **162**, 1389–1400.

Rivin, C. J., C. A. Cullis, and V. A. Walbot (1986) Evaluating quantitative variation in the genome of *Zea mays*. *Genetics* **113**, 1009–1019.

Rogers, S. O., and A. J. Bendich (1987) Heritability and variability in ribosomal-RNA genes of *Vicia faba*. *Genetics* **117**, 285–295.

SanMiguel, P., A. Tikhonov, Y. K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P. S. Springer, K. J. Edwards, M. Lee, Z. Avramova, and J. L. Bennetzen (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765–768.

Schmidt, R. (2000) Synteny: Recent Advances and Future prospects. Current Opinion in Plant Biology **3**, 97-102

Sears, E. R. (1954) The aneuploids of common wheat. *Mo. Agric. Exp. Stn. Res. Bull.* 572.

Song R. T., V. Llaca, E. Linton, and J. Messing (2001) Sequence, regulation, and evolution of the maize 22-kD alpha zein in gene family. *Genome Res.* **11**, 1817–1825

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant, *Arabidopsis thaliana*. *Nature* **408**, 796–815.

Yuan, Y. N., P. J. Sanmiguel, and J. L. Bennetzen (2002) Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *Zea mays*. *Genome Res*. **12**, 1345–1349.

CHAPTER

2

The Basic Toolbox—Acquiring Functional Genomic Data

Overview

"If the only tool you have is a hammer, then everything looks like a nail."

This old adage really does apply to many scientific situations and has shaped the historical investigations of plant form and function. When the tools were ruler and microscope, growth studies and detailed structural descriptions were all that were possible. As the molecular technology developed both the range of studies and the way that questions can be framed have been greatly expanded. As the technology improves old questions can be revisited and new explanations can be suggested.

The new tools that are available for investigating gene structure and function have been steadily developed over the past 30 years. The molecular biology revolution for the characterization of genomes began with the development of recombinant DNA techniques. Today the molecular tools include various cloning vectors, the incorporation of robotics into high-throughput methodologies, for example, in the area of DNA sequencing, and mass spectroscopy for the detailed characterization of proteins. The application of these methodologies results in the generation of very large amounts of data that need to be processed. Whereas in the past the actual accumulation of the data was the rate-limiting step, the bottleneck is now the ability to analyze all the data.

The wealth of data generated by high-throughput methodologies will advance our understanding of gene structure and function by the molecu-

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

lar characterization of already existing variants. In addition, the ability to change gene expression in vivo, by using insertional mutagenesis, RNA interference, or other silencing mechanisms, will be crucial in determining the specific function of a particular gene. Therefore, at the present time, techniques are available to identify gene expression at various stages of development and/or in response to biotic or abiotic stresses and then to develop the biological material to determine which of these observations or structural entities are causal of the changes seen and which are simply the downstream result of some earlier modulation of gene expression.

This chapter considers the various techniques used in the acquisition of genomic data. Broadly speaking, they cover the following main areas:

1. Methods of isolating and fractionating genomes into manageable-sized pieces, with the associated automation and tracking systems that are necessary to manage the experiments and to interpret the results. Genome fractionation must occur at both the DNA and RNA levels so that the actual expressed genomic regions can be determined. The cloning of both genomic DNAs and expressed RNAs is therefore necessary.
2. The development of microarray technology has opened up the possibilities of expression profiling, the visualization of the expression of many genes simultaneously.
3. The downstream processing of the RNAs into proteins and the modification of these proteins and their abundances can also be determined so that the effective contribution of any expressed RNAs can be more directly demonstrated. The development of metabolic profiling will continue to open up new avenues for understanding the function and contribution each of these proteins to the phenotype.
4. The informatics tools to analyze this wealth of molecular data.
5. The ability to select a particular gene or suite of genes and to selectively interfere with their expression, to directly test whether the conclusions drawn from the molecular data actually hold up in practice.

Cloning Systems

The primary problem of fractionating the genome into manageable bits was basically solved with the advent of cloning. This methodology, whatever the vector system used, results in a collection of large numbers of separable fragments. Subsequently, the collection must be screened or additionally characterized to identify the fragments that are of interest. Much of the current development of new vectors and kits has been done by biotechnology com-

panies, and the data and protocols are available from their websites. These developments have made the cloning of both DNA and RNA more routine.

Plasmid-Based Vectors

Most of these cloning vectors are well described and are available in various forms from the various biotechnology companies. The many plasmid-based vectors that are available have been engineered for specific tasks, either for the sequencing or for the expression of the inserted fragment. Included in this set of specialized vectors are those that also add a short peptide sequence to the open reading frame to enable protein-protein interactions to be characterized, an example being the yeast two-hybrid systems (more fully described in Chapter 6) (Bendixen et al., 1994). There are still many uses for plasmid cloning systems including the generation of small fragments of DNA for sequencing, the isolation of cDNAs, and especially full-length cDNAs, and for expressing genes in heterologous systems. The main limitation of plasmid-based systems is the small size of the insert that can be accommodated.

One of the more time-consuming processes is the subcloning or shuttling of fragments of DNA between different vectors. One of the technologies that have been developed to facilitate these rearrangements is the Gateway™ Technology from Invitrogen. Gateway™ Technology is a universal system for cloning and subcloning DNA sequences, facilitating functional gene analysis and protein expression. Gateway™ Technology enables the rapid cloning of one or more genes into virtually any protein expression system. This *in vitro* technology greatly simplifies the process of gene cloning and subcloning. As genes are shuttled between expression vectors both the correct orientation and reading frame are maintained. Gateway™ uses site-specific recombination, effectively eliminating the requirement to work with restriction enzymes and ligase after the initial entry clone is constructed. Once the entry clone has been constructed, the gene of interest can be transferred into a variety of Gateway™-adapted expression vectors (destination vectors). Because the reading frame and orientation of the DNA fragment are maintained during recombination, the new expression clone does not need to be sequenced with each new construct. Two reactions, BP and LR, constitute the Gateway™ Technology (Figure 2.1). The BP reaction uses a recombination reaction between an *att*B DNA segment or expression clone and an *att*P donor vector to create an entry clone. The LR reaction is a recombination between an *att*L entry clone and an *att*R destination vector. The LR reaction is used to move the sequence of interest to one or more destination vectors in parallel reactions. Constructing a Gateway™ expression clone is accomplished in just two steps:

1. The gene of interest is cloned into an entry vector via PCR or traditional cloning methods.

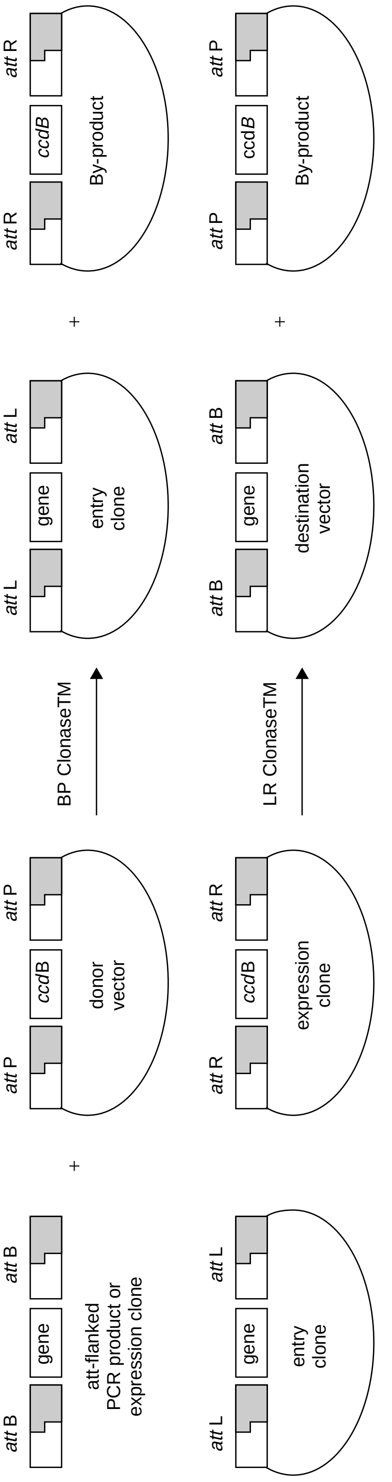


FIGURE 2.1. The reactions involved in the Gateway™ Technology for moving DNA sequences between vectors (Copyright 2003 Invitrogen Corporation. All Rights Reserved. Used With Permission).

2. The entry clone containing the gene of interest is mixed with the appropriate destination vector and Gateway™ LR Clonase™ enzyme mix to generate an expression clone.

Large-Insert Vectors

Three types of vectors that can accommodate large inserts are based on one of bacteriophage λ, yeast artificial chromosomes (YAC) (Kusumi et al., 1993), and bacterial artificial chromosomes (BAC) (Peterson et al., 2000). Each of these has its own particular advantages and disadvantages. The λ-based vectors are relatively easy to screen but pose problems in the subsequent manipulation of each specific recombinant. YACs can accommodate the largest inserts, but the libraries are difficult to maintain. The ability to store the BAC clones frozen and to apply automation to the analysis of these libraries has resulted in a growth in the use of BAC clones. Large-insert libraries can be used for applications such as the construction of a physical map and the map-based cloning of genes. For any of the vectors it is important to ensure that a sufficiently large and representative library is constructed, so that there is a high probability that the particular region of interest is present in the library. The probability of finding a single copy sequence in a library is given by:

$$N = \ln(1-P)/\ln(1-I/C)$$

N is the number of clones generated
P is the required probability that the sequence is present
I is the average insert size in base pairs
C is the genome size in base pairs

The number of clones that need to be screened to find a single copy sequence in libraries that were constructed from DNAs of various plant species is given in Table 2.1.

Table 2.1. The number of clones necessary for having a 99.995% probability (10X coverage) of finding a single copy sequence in a library with 100,000 bp inserts for various plant species

Species	Genome size (Mbp)	Number of clones required
Arabidopsis thaliana	0.125	12,500
Maize	2.73	273,000
Wheat	17	1,700,000
Fritellaria assyrica	124	12,400,000

Bacterial Artificial Chromosome Libraries

These are now a staple resource in the plant genomics community. The libraries can be maintained at low temperatures and are very easily adapted for use in high-throughput automated processes. The insert size that can be accommodated is sufficiently large to generate a manageable library for most plant genomes. The clones can be picked and stored in 96- or 384-well plates for use with most liquid handling systems. Automated procedures for the isolation of BAC DNA followed by the fingerprinting of these clones (see Chapter 3) have made such libraries the material with which most physical maps are generated. Two different vectors are available for making BAC libraries. These are the standard bacterial artificial chromosome (BAC) vector, and the binary BAC (BIBAC) vector. The BIBAC vector is based on the standard BAC vector for genomic libraries, with the addition of regions from the binary vector system for *Agrobacterium*-mediated plant transformation (http://hbz7.tamu.edu/homelinks/tool/bac_content.htm; http://www.research.cornell.edu/Biotech/BIBAC/BIBACHomePage.html). This provides the opportunity for the direct transfer of the recombinants to *A. tumefaciens* and subsequent use for plant transformation. One of the possible drawbacks of the BIBAC vector is that its larger size may interfere with the automated DNA fingerprinting processes because of the large number of overlapping bands generated from the larger vector DNA. The vectors are in constant modification, and a vector that includes the best features of both the P1 and BAC systems is shown in Figure 2.2.

Generation of BAC Libraries

The source DNA for BAC libraries can be isolated in at least two ways. Extractions can be made from whole cells, in which case the organellar genomes will comprise a substantial fraction of the clones. Alternatively, nuclei can be isolated and then the DNA purified from these nuclei. In this latter case any organellar sequences identified in the library are likely to come from copies that have been integrated into the nuclear genome. The steps involved in preparing the BAC library include:

- The megabase-sized DNA is isolated from cells or nuclei.
- The DNA is then embedded in agarose plugs and partially digested by the restriction enzymes of choice.
- The size range of 100–350 kb from the partially digested DNA is selected after separation of the partial digest on a gel.
- A second size selection can be performed if required to eliminate small trapped fragments from the first gel run.
- The size-selected DNA is then ligated with a BAC vector of choice, the latter having been first digested with the appropriate enzyme and

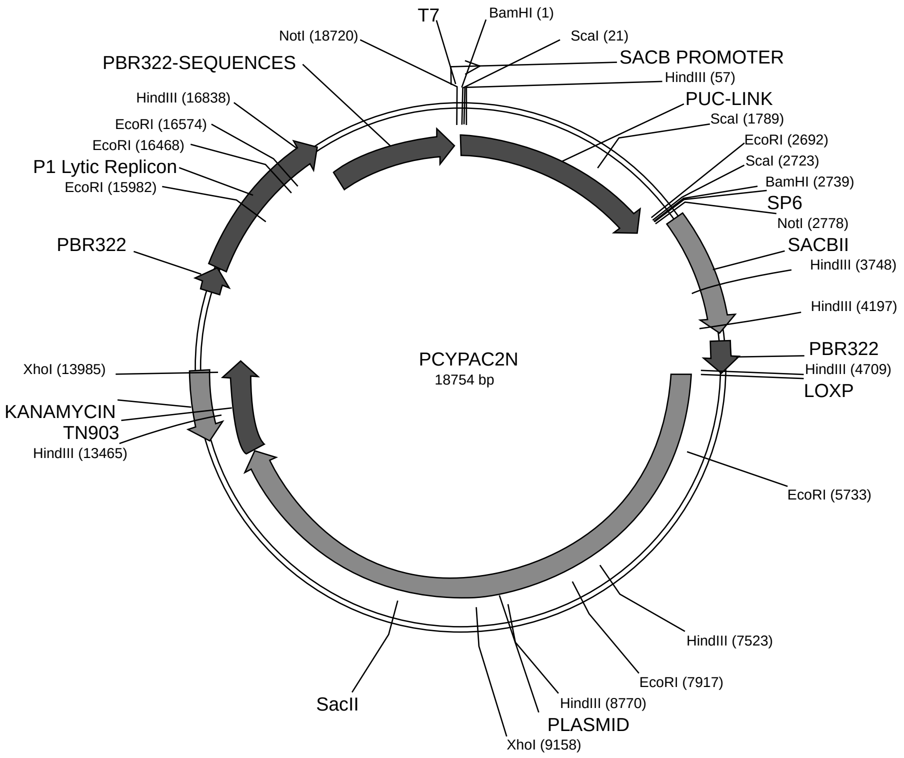


FIGURE 2.2. Map of the PAC/BAC vector PCYPAC2N from http://www.chori.org/bacpac/pcypac2.htm.

then dephosphorylated; the ligation mixture is then electroporated into the appropriate *Escherichia coli* host strain.

- BAC transformants are then usually selected on LB plates containing an antibiotic, X-Gal, and IPTG.
- White recombinant colonies are picked robotically and stored as individual clones in 96- or 384-well microtiter plates as glycerol stocks at –80°C.
- The library can then be replicated to provide working copies and a master (original) copy.
- Before extensive use, the library should be evaluated for at least three quality factors:
  - Insert size distribution
  - Chloroplast and mitochondrial DNA content
  - Genome representation as determined by screening the library with single-copy markers that are dispersed throughout the genome.

Full protocols for generating such libraries are available from various sources (Peterson et al., 2000; http://hbz7.tamu.edu/homelinks/tool/bac_content.htm; http://www.research.cornell.edu/Biotech/BIBAC/BIBACHomePage.html).

Utilization of BAC Libraries

The BAC libraries can have a number of different uses.

- One increasingly popular use is to fingerprint large numbers of BACs to develop physical maps. This is described in detail in Chapter 3.
- Libraries can also be used to isolate "similar" (potentially syntenic) regions across species. To do this, the library must be screened with sets of specifically designed probes. The probes are usually either cDNAs or oligonucleotides that have been conserved across species, the latter termed "overgo oligos." A program available at http://www.mousegenome.bcm.tmc.edu/webovergo/OvergoDescription.asp can be used for designing overgo probes. Relatively nonredundant sequences are usually selected for the design of a pair of overlapping oligos (overgos) for each sequence, either from expressed sequence tag (EST) sequences or BAC end sequences. The screenings are performed by hybridization, usually to high-density nylon filters onto which a large number of BAC clones have been spotted. The clones are usually double spotted to avoid false positives, because a significant amount of work is involved in analyzing each positive BAC clone. Overgos (in a microtiter plate, for example) can be pooled and the BAC filters hybridized with the labeled pool. The BACs that hybridize to the pooled overgo probes can be rearrayed and hybridized with columns and rows from the original pooled plate. In this way, specific BACs that hybridize to specific overgo probes can be identified in just $n + m + 1$ hybridizations (where n and m are the number of rows and columns in the original plate). Thus all 384 oligos in a 24×16-well plate can be unambiguously assigned to specific BACs with just 41 hybridization reactions (see Figure 2.3).
- This approach can be used to test synteny across species. The syntenic approach involves the selection of BAC clones either containing a known gene or located at particular positions on a specific chromosome. The sequence comparisons of the chosen BACs and the chromosomes will provide evidence for the conservation, or otherwise, of these regions across ever-increasing evolutionary distance. Obviously, for this approach to be viable it is necessary to have a large BAC library for each of the species under investigation because syntenic regions, almost by definition, must contain low, or single-

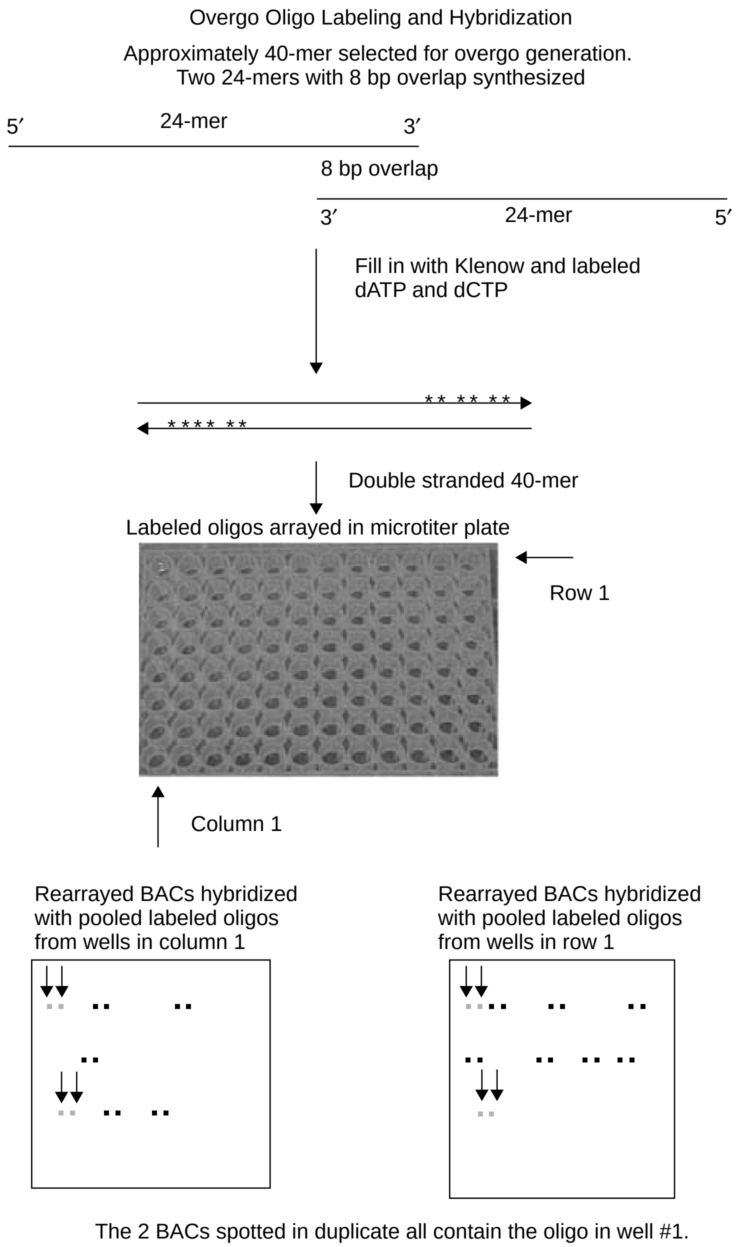


FIGURE 2.3. BAC screening with overgo oligo probes. The overgo probes are labeled and pooled. Each pool is then hybridized to a high-density filter of BAC clones to identify which of the BAC clone(s) contain each of the overgo oligos.

copy, sequences for their initial selection. The development and public availability of BAC libraries for many species are currently being funded, and strategic decisions must be taken as to which would be the most informative libraries to have (see http://ucjeps.berkeley.edu/bryolab/greenplantpage.html).

- BACs containing a high density of genes can be selected by using hybridization with expressed sequences to determine the gene richness of the region. This is certainly one way of trying to identify gene-rich regions of the genome, as a prelude to sequencing the "important" regions of large complex plant genomes (see Chapter 3).

One of the decisions to be made in the development of any BAC resource is the identity of the particular plant or line that will be the source of the DNA from which the BAC library will be prepared. In many cases this is a community (i.e., those interested in the particular plant species or family) decision as to which would be the most appropriate starting point. Once the initial libraries have been prepared it will be easier to use the information garnered from these libraries to survey a much larger diverse population to understand the amount of variation that is available. An example of this process in action is the choice of the two initial varieties for the preparation of BAC libraries from banana (*Musa*). The Global *Musa* Genomics Consortium, a publicly funded group of 27 collaborating organizations from 13 countries, was launched in July 2001. As a result of the deliberations of this group two banana BAC libraries, one from *M. acuminata* "Calcutta 4" and the other from *M. balbisiana* "Pisang Klutuk Wulung," have been constructed and made available to the Consortium members as the standard cultivars for characterization. Because banana is a monocot it is likely that the information from the rice genome would be generally applicable to gene discovery in this organism. Therefore, the expectation is that these BAC libraries will enable the wealth of information from the rice genome sequence to be leveraged for the ultimate benefit of smallholder farmers throughout the tropics.

cDNA Cloning

The term *cDNA* is short for complementary DNA, because messenger RNA (mRNA) cannot be cloned directly but a DNA copy of the mRNA can be cloned. The conversion of mRNA to DNA is accomplished by the action of reverse transcriptase and DNA polymerase (Gubler and Hoffman, 1983). The reverse transcriptase makes a single-stranded DNA complement of the mRNA. The second DNA strand is generated by DNA polymerase. The double-stranded product can be introduced into an appropriate plasmid or lambda vector. The availability of high-throughput sequencing methods opened the way to developing EST collections from cDNAs. ESTs are short fragments of cDNAs that are usually biased toward the 3′ end of the mRNA.

To generate ESTs it is not necessary to generate full-length sequences. However, full-length cDNA clones can greatly facilitate both gene annotation and functional studies and are an important resource. A scheme to isolate full-length cDNA clones is given in Chapter 4.

The identification of the mRNA population in a given cell or tissue is necessary to gain an understanding of gene expression in that cell or tissue. Another important function is to compare mRNA populations when the plant material has been subjected to a variety of perturbations, such as the stress of pathogen invasion, to characterize the changes in gene expression under these altered growth conditions. The analysis of differential gene expression is covered in detail in Chapter 6.

The mRNA can be fractionated before cloning to overcome the wide range of messenger abundances within the mRNA population by the normalization and subtraction of libraries. Alternatively, the mRNAs can be characterized by using various lengths of "tags" that unambiguously identify a particular message and that also quantify the mRNA levels. These methods include serial analysis of gene expression (SAGE) (Powell, 1998; Madden et al., 2000) and massive parallel sample sequencing (MPSS™) (Brenner et al., 2000), which are discussed in greater detail in Chapter 6. Alternatively, the mRNA differences can be visualized with differential display methods.

SUBTRACTION LIBRARIES

The individual mRNAs that are isolated from a tissue can be present in that population over many orders of magnitude. To prevent the excess redundant characterization of the same message, normalization or subtraction is used to reduce the disparity of representation within these populations. The construction of these modified libraries takes advantage of some combination of hybridization kinetics and amplification. For normalization, two cDNA populations are hybridized to reduce the abundance of the most prevalent messages. The remaining cDNAs are then cloned and either sequenced or used in some other expression profiling experiment (Figure 2.4). Alternatively, suppressive subtracted hybridization (SSH) (Diatchenko et al., 1998; Figure 2.5) involves the hybridization and amplification of the cDNAs to physically remove the highly abundant sequences. This type of reduction in the representation of the highly abundant messages in the mRNA population is particularly necessary to identify rare messages. The normalization of a library should result in all the sequences, whether originally abundant or rare, being present in about the same frequency after normalization. SSH, on the other hand, will result in the selective representation of sequences that are only in one of the cDNA libraries. The efficacy of these methods must be weighed against the costs of redundant sequencing of large numbers of ESTs. Because none of the subtraction/normalization methods

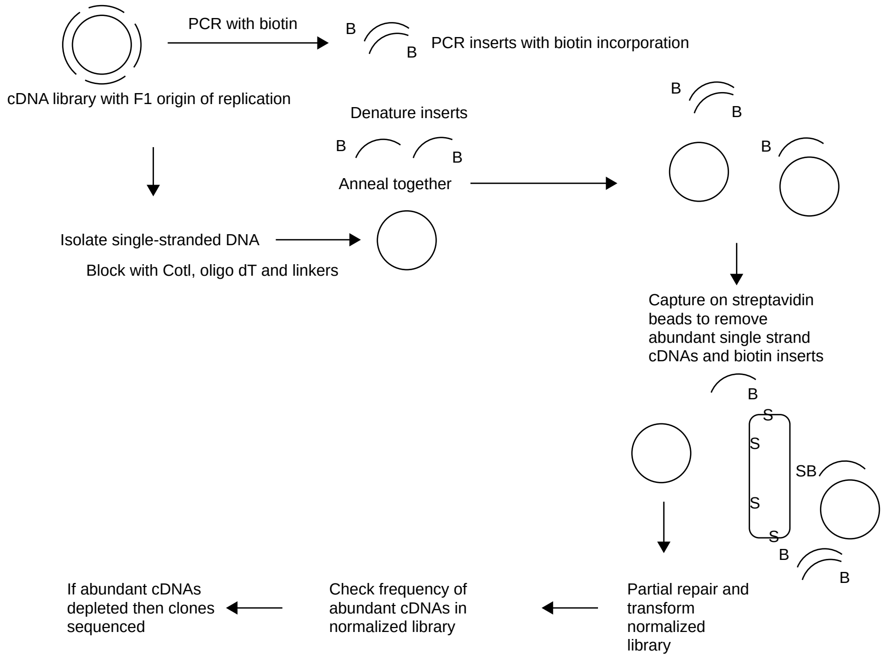
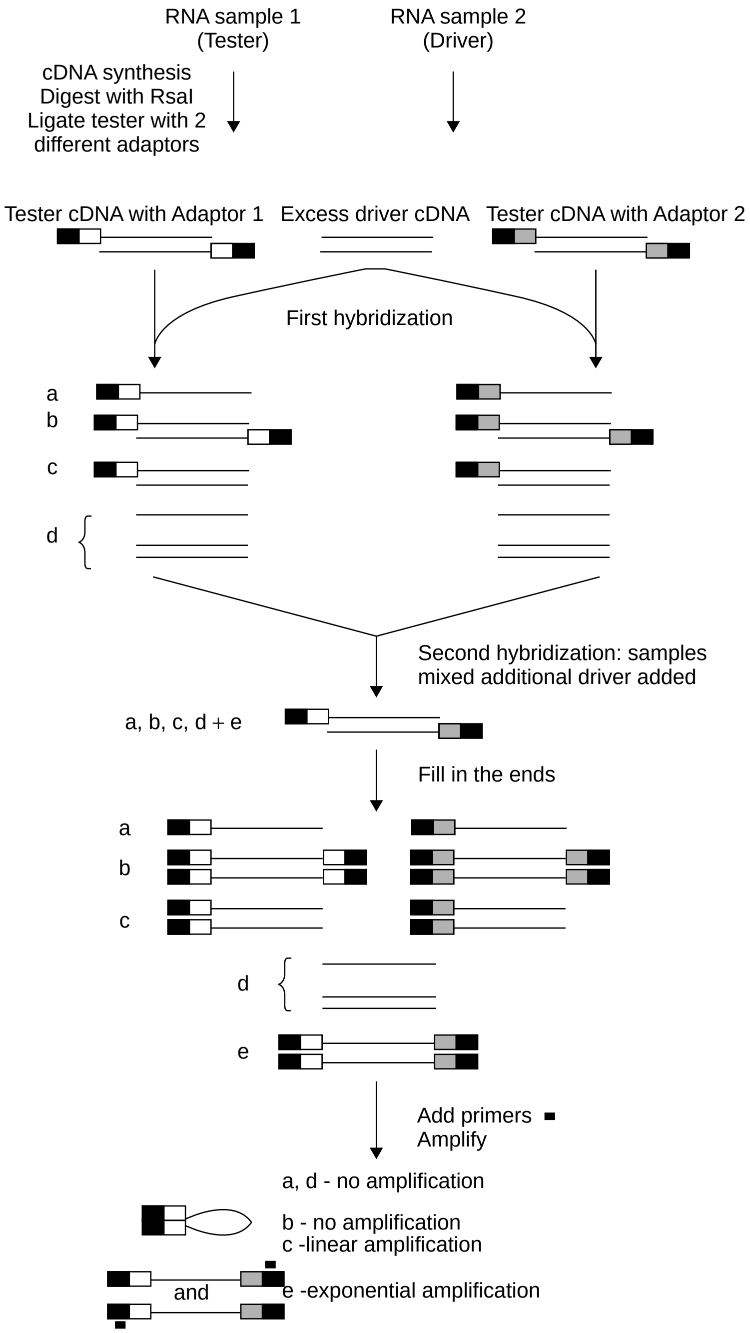


FIGURE 2.4. Normalization of cDNA libraries. The cDNA library is divided into two sets, one in which the inserts are amplified with the incorporation of biotin, and the other being the isolation of single-stranded circles. The two are appropriately annealed together, and the resulting mixture is bound to streptavidin beads. The unbound circles (low-abundance cDNAs in the library) are repaired and retransformed. The normalized library is then sequenced. (Adapted from http://hg.wustl.edu/COGENE/INFO/norm.html).

FIGURE 2.5. SSH scheme (from http://www.clontech.com/products/literature/pdf/brochures/PCR-SelectBR.pdf). The 2 samples to be subtracted are isolated and converted to cDNA. The tester sample is divided into 2, and each fraction ligated to a specific adaptor. The tester is then hybridized with excess driver cDNA that did not have adaptor added. After the first hybridization the samples are mixed, additional driver is added, and the mixture is hybridized again. The ends of the molecules are then filled in and amplified by using the adaptor sequences as primers. Only those messages that are rare in the tester sample are amplified. These can then be cloned and sequenced. (Reprinted with permission from BD Biosciences Clontech.)


RNA sample 1
(Tester)
RNA sample 2
(Driver)
cDNA synthesis
Digest with RsaI
Ligate tester with 2
different adaptors
Tester cDNA with Adaptor 1
Excess driver cDNA
Tester cDNA with Adaptor 2
First hybridization
a
b
c
d
Second hybridization: samples
mixed additional driver added
a, b, c, d + e
Fill in the ends
a
b
c
d
e
Add primers
Amplify
a, d - no amplification
b - no amplification
c -linear amplification
and
e -exponential amplification

completely remove the abundant messages, there will still be a significant waste in sequencing the subtracted libraries as the common messages will already have been identified.

Differential Display

Differential display technology works by the amplification of the 3′ terminal portions of mRNAs and the visualization of those fragments on a DNA sequencing gel (Liang and Pardee, 1992). The anchored oligo-dT primers define the 3′ end of the RNA, and then a limited number of short arbitrary primers are used to amplify most of the mRNA in a cell (Figure 2.6). The separation of the amplified fragments by denaturing polyacrylamide gel electrophoresis allows direct side-by-side comparison of most of

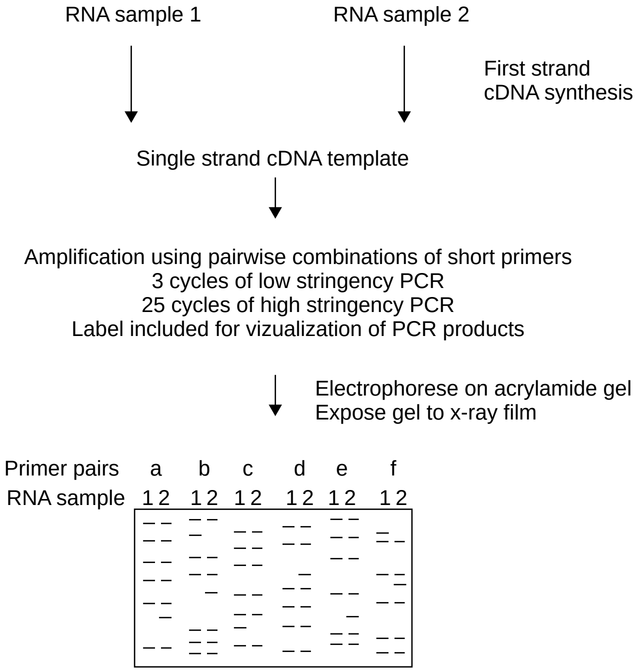


FIGURE 2.6. Differential display. Each sample is treated in the same way. The first-strand cDNA is synthesized. This is then amplified with a series of short primers for 3 cycles at low stringency to account for the length of the primers. After the first 3 cycles the amplification is continued for a further 25 cycles at high stringency. In the amplification, label is included so that the products can be visualized. The products are then separated on a polyacrylamide gel, and the bands are compared. Differential bands can be excised and characterized further.

the mRNAs. The polymorphic bands representing the differentially expressed sequences can be excised from the gel, reamplified, cloned, and sequenced.

DNA Microarrays

The combination of sequence information and automation has facilitated the change from looking at one gene at a time to being able to profile expression and changes of expression for many genes. DNA microarrays, which are one of the basic formats for looking at these changes, are orderly arrangements of DNA samples fabricated by high-speed robotics on glass, nylon, or other substrates. The initial experiments used cDNA immobilized on the surface with robot spotting, although a current alternative to the use of PCR products is oligonucleotides either synthesized on the chip or synthesized off the chip and anchored later. These arrays can be used to detect polymorphisms and mutations, as well as to help map genomic libraries and characterize gene expression.

The schema for a microarray experiment is shown in Figure 6.2. The steps involved are:

1. Development of the microaray. As mentioned above, this can be an array based on either fragments amplified through PCR or small oligonucleotides that are attached to the substrate. In either case, the sequences to be placed on the microarray must be selected. The PCR products, for example, can be amplified from cDNAs that comprise a unigene set developed from EST collections (see Chapter 4 and Figure 4.1) or from genomic DNA by using primers based on gene predictions from a genomic sequence. The oligonucleotides can be designed from the same information source.
2. The arrays are then printed by using the information. When the PCR products or the oligonucleotides are first generated and then attached to a substrate, this results in a more flexible platform because the array designed can be changed relatively easily and less expensively than when the oligonucleotides are actually synthesized on the substrate.
3. The probes are labeled and hybridized to the microarrays. The design of the experiment and the number of replicates are important parameters for any differences observed to be statistically validated.
4. The hybridization data are analyzed with a series of software programs, and any patterns of coordinated changes in gene expression are detected.
5. The form in which the data are reported is important for comparison between experiments. A series of standards for such reporting have

been developed, and are continuing to be developed, and are known as the minimal information about a microarray experiment (MIAME; Brazma et al., 2001).

Sequencing Strategies and Automation

The development of automated sequencing strategies and the reduction in the cost of sequencing have allowed the design of experiments that rely heavily on the acquisition of nucleic acid sequence information. This reduction in cost has also meant that redundant sequencing is no longer necessarily a bar to identifying the sequence of interest. Therefore, shotgun sequencing strategies that result in the same sequence being generated many times to ensure a complete sequence assembly is now a viable option. The complete sequence of a BAC can be obtained by the sequencing of subcloned random fragments that have been isolated from sheared DNA. These sequences are then assembled electronically to generate the linear sequence of the BAC clone. Obviously, one potential drawback or problem in this approach would be the presence of multiple copies of identical repeats within the BAC, longer that the normal sequence read, that would be difficult to assemble. These gaps can be filled by using other strategies if necessary. At least two approaches are available. In the first approach, regions of the shotgun reads that are unique, and not close to the end of the read, can be used to generate primers for additional sequencing reactions. This enables the sequence to be read through repetitive regions provided they are short enough to be covered in a single sequencing run. With improvements in technology this distance is now extending to more than a kilobase in length. The second approach is to saturate the target sequence with a transposon and use a primer close to the edge of this transposon to generate sequence reads. This latter approach can get out of repetitive regions because the read can start in the repetitive sequence and, provided the insertion is positioned sufficiently close to the end of that region, read into the adjacent low-copy-number sequence. Although the cost of generating the DNA sequence may no longer be a limiting factor for almost any project, informatics challenges associated with this rapid generation of sequence data are likely to provide significant bottlenecks in data analysis for the foreseeable future.

Protein Characterization

In many ways the approach to generating and understanding the protein sequence has come full circle. Originally it was thought that the way to identify the sequences of genes would be by sequencing the proteins, and Sanger was awarded the Nobel Prize (1958) for devising methods of getting peptide

sequences. However, it then turned out that DNA sequencing was much faster, leading to the explosion of DNA sequence data (and another Nobel Prize for Sanger in 1980), with the protein data coming from the nucleic acid sequence (usually from the cDNA sequence because, even now, gene prediction programs still need much development to be applied with confidence). The application of mass spectrometry to protein characterization and identification has improved the amount of information that can be obtained from biological samples, including the ability to do peptide sequencing on very small samples. The proteome can be described as the full protein complement of an organism, cell type, or tissue at any one moment in time. Proteomic studies are still in their infancy but are becoming more important and pervasive. One key technology for proteome analysis is two-dimensional gel electrophoresis (2-DE), which has the capability to separate very complex protein mixtures. These mixtures can contain thousands of components with different physicochemical properties and abundances. A second key technology is mass spectrometry, which is useful for proteome analysis because it has a sensitive detection range and can be used for high-throughput identification. The results obtained from these techniques can be applied to gene and protein database searches and can also be used to identify posttranslational modifications of proteins (information that is impossible to acquire from nucleic acid studies). The most important ionization techniques for proteomics, matrix-assisted laser desorption/ionization (MALDI) mass spectrometry (MS), and electrospray ionization (ESI), have been continuously improved so that MS plus MS/MS data have enhanced the protein identification capabilities of the methods (Figure 2.7) (Pandey and Mann, 2000).

Proteome analysis is now a complementary and potentially coupled technology to transcription profiling. However, proteome-based studies must be carefully designed and performed to ensure reproducible analysis. In particular, the selection of the tissue and its preparation are crucial steps in proteome analysis.

Databases and Informatics

The high-throughput methodologies that have been developed for both DNA and proteomics have highlighted the need for sophisticated informatics to deal with the data. Additionally, it is also essential to define the appropriate structures for the databases that house all this information so that the data can be accessible both immediately and in an archival form.

Databases

The creation of biological databases represents a fundamental change in the way biological information is disseminated. Future advances in biology will

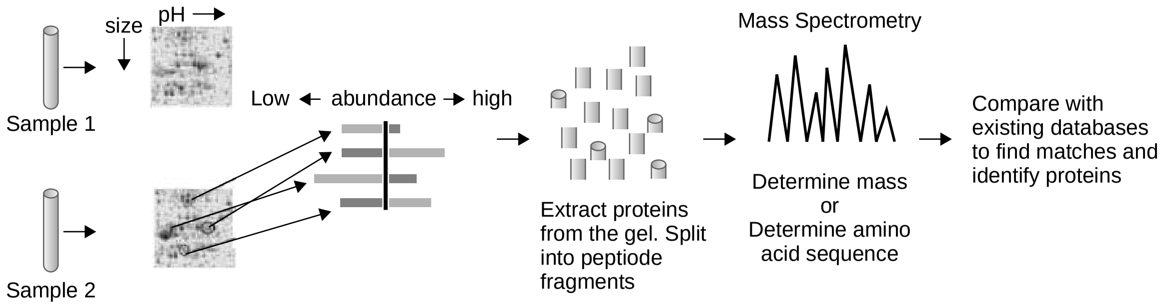


FIGURE 2.7. Proteomics experimental flow. The isolation of proteins followed by their separation by 2-D gel electrophoresis. The patterns of proteins are compared, and those of interest are excised, fragmented, and separated by mass spectrometry. The amino acid composition or sequence is determined (depending on the particular MS technology applied), and the databases are searched to identify the protein. The actual amino acid sequence is important for those species for which there is little genomic or cDNA sequence. ▢ Protein abundance from sample 1; ▩ protein abundance from sample 2.

depend, in large part, on the improvement of critical databases. However, organism-specific researchers want to generate and use data locally, annotate it as needed, and answer very specific questions driven by physical experimentation, even though these experimental data and findings often are shared with the global community. This larger group of researchers, namely the global community, requires access to large amounts of data to address questions that might be of limited interest to its original producers. Furthermore, the tools and data formats applicable to this high-level sharing may be very different from those applied in the context of the original data production. The underlying challenge, therefore, is to integrate diverse data and data types in order to provide a more complete and consistent view of the information contained therein and to provide the means to increase the utility of these resources as the quantity of data increases in the future.

Currently, however, the distributed data resources of biology, and in particular of plant genomics, share a number of characteristics that have made the actual interrogation and analysis problematic. In many instances the data are in flat file structures and there is no separate schema for the metadata or such a schema is not available. The data are addressed through varying call-based interfaces, rather than through a declarative query language. This disallows the use of agents and requires human-computer interactions. Indeed, the data collections that result from the genomics revolution represent a need to change the way in which biological data are disseminated.

All databases should allow the broadest access possible to accommodate these altered data usages. Therefore, a standard format must be developed to facilitate this access. Included in the consideration of such formats would be that the associated metadata must also be accessible to enable queries by machine agents as well as individuals. Wherever possible, the development and use of organism- and/or database-independent software should be encouraged. This will not diminish the importance of organism-dependent repositories but will enable the data to be used more widely and will increase their importance.

The essential information associated with these data repositories, or particular entries within the databases, should also be reported. For example, a list of possible information associated with comparative genetic mapping in plants could include the raw segregation data for the individual mapping populations, explicit criteria that were used to determine whether two markers represent orthologous loci, and the sequences of all DNA-based markers as GenBank files. Trace files for sequence-based polymorphisms should be archived by mapping laboratories for future access, and the underlying information used to construct the physical maps (e.g., FPC fingerprints, BAC end sequences, BAC hybridization) could be made publicly available in a project database. The structure of the databases and the information contained therein must be supported by appropriate documentation and

standard operating procedures for both experimentation and analysis. A set of criteria somewhat like those developed for microarrays, but perhaps in the form of the minimal information associated with a functional genomics experiment (MIAFGE) would go some way to solving this problem. A start on generating such metrics has been made by the Plant Genome Research Program (http://plantgenome.sdsc.edu/Awardees Meeting/Bioinformatics_and_Databases/).

Given a collective effort, the bioinformatics landscape could be transformed from a small number of insular database projects to a large number of open, interoperable data services that together would form the fabric of a new biological data infrastructure. This transformation would require a change of emphasis, reducing the efforts on species-specific databases and redeploying them to the development of biological data service infrastructure from which deliverables would be portable, general-purpose software and standards that would be made freely available to the academic community as well as industry. Databases such as ZMDB and Gramene that are developing both the tools and open structures to allow the broadest access are examples of these efforts.

Informatics Tools

A range of informatics tools is needed to analyze this massive generation of data. For nucleic acids, these tools range from those dealing with sequence quality and assembly to search engines for comparing sequences and also include programs that can be used to annotate DNA sequences that are trained to identify genes.

Sequence Quality

Generally the sequence quality is checked by the length of the reads and quality score assignment. Phred reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files (Ewing et al., 1998). These files can be in any one of three formats: FASTA, PHD or the SCF. Quality values for the bases are written to both FASTA and PHD format files. These files can be used by the Phrap sequence assembly program to increase the accuracy of the assembled sequence.

Phrap is a leading program for DNA sequence assembly. Phrap is used to locate overlapping regions within individual sequences and assemble them into longer contiguous sequences (contigs). Phrap is most commonly used for assembling data from shotgun sequencing but can also be used for EST clustering, genotyping, and identifying sequence polymorphisms. Phrap uses Phred's quality scores to determine a highly accurate consensus

sequence by examining all the individual sequences at a given position. This approach is especially important in regions of low coverage or regions of systematic errors. The quality of the consensus sequence is also estimated from the quality information of individual sequences.

A more common program for EST clustering is CLUSTAL W (Thompson, 1994). This is a freely available and portable program for multiple sequence alignment. Because EST projects are essentially one-pass sequencing of cDNAs it is important to cluster the derived sequences into contiguous sets that come from the same gene or from members of a gene family. This will result in a consensus sequence of that transcript. The consensus sequence can be used, for example, to design overgo oligos for screening related BAC libraries and for designing primers to test whether nucleotide polymorphisms within the clustered sequence are the result of single-nucleotide polymorphisms (SNPs) or sequencing errors. The clustered consensus can also be longer than any of the individual ESTs, thereby extending the length of the known transcribed sequence.

As well as needing to cluster ESTs, shotgun reads of genomic sequences also must be assembled. The program CAP4 (Huang et al., 2000) utilizes base quality values, forward-reverse constraints, and automatic clipping of poor-quality ends based on overlaps to assist in assembly and production of more accurate contigs. CAP4 generates contigs and consensus sequences that can be viewed and edited with Paracel's AssemblyView, the University of Washington's Consed (Gordon et al., 1998), or Staden's gap4 Contig Editor. Furthermore, CAP4 also generates valuable information concerning scaffolds, that is, what contigs are linked together based on constraints. This feature is especially important for low-pass sequencing projects to order the contigs and for finishing phases by providing information on which subclones are necessary to bridge the gaps.

The whole suite of informatics resources and needs are described in more detail in Chapter 9.

SUMMARY

The technology for acquiring molecular data about the whole plant kingdom has progressed, and continues to progress, rapidly. These technologies, and the resources being generated by large infrastructure projects, should allow the whole community of plant biologists to gain new insights into the workings of their favorite organisms. The volume of data being generated is almost overwhelming, and there is a need to ensure that these data remain available as projects are completed. In addition, the need for informatics will continue to grow, both in the realm of software to visualize the data and for the analysis of the data.

References

Andersen, J. S., and M. Mann (2000) Functional genomics by mass spectrometry. *FEBS Lett.* **480**, 25–31.

Bendixen, C., S. Gangloff. and R. Rothstein (1994) A yeast mating-selection scheme for detection of protein-protein interactions. *Nucleic Acids Res.* **22**, 1778–1779.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. P. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29**, 365–371.

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays *Nat. Biotechnol.* **18**, 630–634.

Diatchenko, L., A. Chenchik, and P. D. Siebert (1998) Suppression subtractive hybridization: a method for generating subtracted cDNA libraries starting from poly(A+) or total RNA. In: *RT-PCR Methods for Gene Cloning and Analysis*. Eds. Siebert, P. and Larrick, J.: BioTechniques Books, Westborough, Massachusetts, 213–239.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185.

Ewing, B., and P. Green (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 86–194.

Gordon, D., C. Abajian, and P. Green (1998) Consed—a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202

Gubler, U., and B. J. Hoffman (1983) A simple and very efficient method for generating cDNA libraries. *Gene* **25**, 263–269.

Huang, X., G. Herrmannsfeldt, T. Jones, J. Qian, S. L. Rash, C. P. Smith, and C. Boysen (2000) CAP4—Paracel's DNASequence Assembly Program http://www.paracel.com/publications/cap4_092200.pdf

Huang, X., and A. Madan (1999) CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877.

Kusumi K., J. Smith, J. Segre, D. Koos, and E. Lander (1993) Construction of a large-insert yeast artificial chromosome library of the mouse genome. *Mamm. Genome* **4**, 391–392.

Liang, P., and A. B. Pardee (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967–971.

Madden, S. L., C. J. Wang, and G. Landes (2000) Serial analysis of gene expression: From gene discovery to target identification. *Drug Discovery Today* **5**, 415–425.

Mann, M., and A. Pandey (2001) Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem. Sci.* **26**, 54–61.

Peterson D., J. Tomkins, D. Frisch, R. Wing. and A. Paterson (2000) Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. *J. Agric. Genomics* vol. 5 (http://www.ncgr.org/jag/papers00/paper300/indexpage300.html).

Pandey, A, and M. Mann (2000) Proteomics to study genes and genomes. *Nature* **405**, 837–846.

Powell, J. (1998) Enhanced concatemer cloning - a modification to the SAGE (Serial Analysis of Gene Expression) technique. *Nucleic Acids Res*. **26**, 3445–3446.

Rowley, A., J. S. Choudhary, M. Marzioch, M. A. Ward, M. Weir, R. C. Solari, and W. P. Blackstock (2000) Applications of protein mass spectrometry in cell biology. *Methods* **20**, 383–397.

Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**, 4673–4680.

Tsugita, A, and M. Kamo (1999) 2-D electrophoresis of plant proteins. *Methods Mol. Biol*. **112**, 95–97.

Web Resources

http://hbz7.tamu.edu/homelinks/tool/bac_content.htm

http://plantgenome.sdsc.edu/AwardeesMeeting/Bioinformatics_and_Databases/

http://ucjeps.berkeley.edu/bryolab/greenplantpage.html

http://www.mousegenome.bcm.tmc.edu/webovergo/OvergoDescription.asp

http://www.research.cornell.edu/Biotech/BIBAC/BIBACHomePage.html

http://www.tree.caltech.edu/protocols/overgo.html

CHAPTER

3

SEQUENCING STRATEGIES

OVERVIEW

The central goal in any genome sequencing project is the discovery of *all* the genes in that organism, which in essence can only be known for certain when every nucleotide in the genome is known and placed in its appropriate context. On December 14, 2000 the first complete genome sequence of a plant was reported in which the analysis of sequenced regions covered 115.4 Mb of a total genome size of 125 Mb (The *Arabidopsis* Genome Initiative, 2000). How was this achieved? Essentially, the *Arabidopsis* genome was sequenced with the traditional approach to genome sequencing. This involved cutting the genome into a large number of subfragments, cloning the pieces, and then grouping the pieces together as large regions of overlapping clones. The chromosomal locations of all of these large groups were determined with the high-density genetic map. Groups of researchers then took on the task of sequencing a specific chromosome or chromosome segment. This sequencing strategy is slow and expensive but provides the most precise and complete sequence across the entire genome. The resulting sequence only represented the inbred progeny of a single isolate of *Arabidopsis*, so any variation between ecotypes was not immediately available.

Complete genome sequences, even those that are made up of a very large number of relatively short segments, of a variety of plants will provide powerful tools for biologists. The sequences will aid in understanding how gene families have been created, amplified, and diverged, resulting in the creation of new biological activities and specificities. The gene content of related species can be investigated and compared to identify which pathways are shared among many species and which are restricted to some parts of the plant kingdom. The advantage of having a full set of genes for a compre-

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

hensive characterization of gene expression especially, for example, when plants are growing under stressed conditions, will lead to the understanding and subsequent manipulation of plant growth for improved agronomic performance. The efficient use of high-throughput approaches like microarray hybridization and analysis will be much more efficient if the complete suite of genes for the plant is available for experimentation. Not only will geneticists have access to the genes underlying quantitative trait loci (QTLs), but they will also be able to generate essentially an unlimited number of DNA markers. One of the major uses of a genome sequence would be to perform map-based cloning of genes and to associate candidate genes with important traits. A well-integrated physical and genetic map is essential for map-based cloning, rendering a pure shotgun sequence less valuable than an anchored sequence. Therefore, the genes identified by sequencing projects will provide the basis for determining why and how some characteristics are shared among particular lineages of organisms while others have a more limited representation in the branches of the tree of life. A complete genome sequence is a first step toward the understanding of biological processes, but it still must be followed by detailed studies of gene function.

The costs and benefits of following a complete genome strategy as opposed to just trying to identify the genes themselves need to be considered. Relatively few plant species have a comprehensive sequence analysis available. Across-species comparisons are very valuable, but how many species are needed to be able to use comparative and syntenic relationships to draw valid conclusions? Therefore, how many genomes need to be sequenced? What are the alternative approaches to getting sufficient information that do not necessarily aim to generate the complete genome sequence but will serve to add enormous value to the existing genome resources?

The traditional approach to genome sequencing (the stepwise sequencing of overlapping clones) is slow and expensive. A faster and less expensive method is the shotgun sequence analysis of small-insert clones (Venter et al., 1996). This latter approach was used to generate the rice genome sequence by Syngenta's Torrey Mesa Research Institute and the Beijing Genomics Institute (Goff et al., 2002). This method, which produced draft sequences, only cost approximately 10% of the International Rice Genome Sequencing Project (ISGRP), which followed the traditional sequencing method. Shotgun sequencing alone does not provide the locations of the sequence segments on the genetic or physical map. In the traditional sequencing approach, however, the positions of the sequenced regions are already known. An alternative to generating a whole genome sequence is to devise a gene enrichment strategy that will result in the sequences of all the genes but without the need to generate the sequences of all the repetitive regions. This gene enrichment strategy will provide a paradigm for a cost-

effective sequencing of large complex genomes of plants that would otherwise be cost prohibitive to produce by whole genome sequence.

In this chapter we consider the processes for generating both whole genome sequences and partial genome sequences enriched for genes. The organization of a whole genome sequence by the traditional and shotgun methods is considered. Various forms of gene enrichment strategies including expressed sequence tags (ESTs), kinetic and methylation-dependent fractionations of the genome, and the use of transposons are explained. The reliance on bioinformatics to assemble such sequences is also considered. A compilation of the "best approach" to obtaining useful genome sequences is described. Finally, the potential plant targets for genome sequencing in the near future and the effect of technology on any predictions are discussed.

Getting a Complete Genome Sequence

Two alternative, but not mutually exclusive, sequencing strategies for the whole genome have been developed (Figure 3.1). The first, as used for deciphering the *Arabidopsis* sequence and the IRGSP rice sequencing effort, is one in which an anchored physical map is generated and then the sequencing is done on all the parts that comprise the physical assembly. The second strategy essentially eschews the need for a prior physical map and generates enormous quantities of random sequence data that are then bioinformatically assembled into a physical genome of some degree of completeness. This second strategy is the shotgun strategy that was used by the Syngenta and Beijing rice genome sequencing projects. It is clearly applicable for a simple genome where every sequence is represented once, or at most a few times, so that a limited number of assemblies can be completed. It has been successfully applied to generate the complete sequence of many microbes. However, the assembly task becomes much more difficult when there are many copies of dispersed repeated sequences, which themselves can be complex. Indeed, there is a difference of opinion as to whether or not the private (Celera) shotgun human sequence could have been assembled into large regions without the publicly available physical map (Butcher, 2001). The problem with plant genomes becomes even more pronounced because the evolutionary history indicates multiple rounds of polyploidy leaving related regions in very different parts of the genome. The highly repetitive nature of the maize genome, for example, which contains large numbers of dispersed very similar repeats, may well be impossible to assemble correctly into a complete genome sequence with the current bioinformatics tools. The assembly of the two rice shotgun sequences each resulted in tens of thousands of contiguous fragments of length less than 7 kb, and the rice genome is only one-fifth the size of the maize genome (Goff et al., 2002).

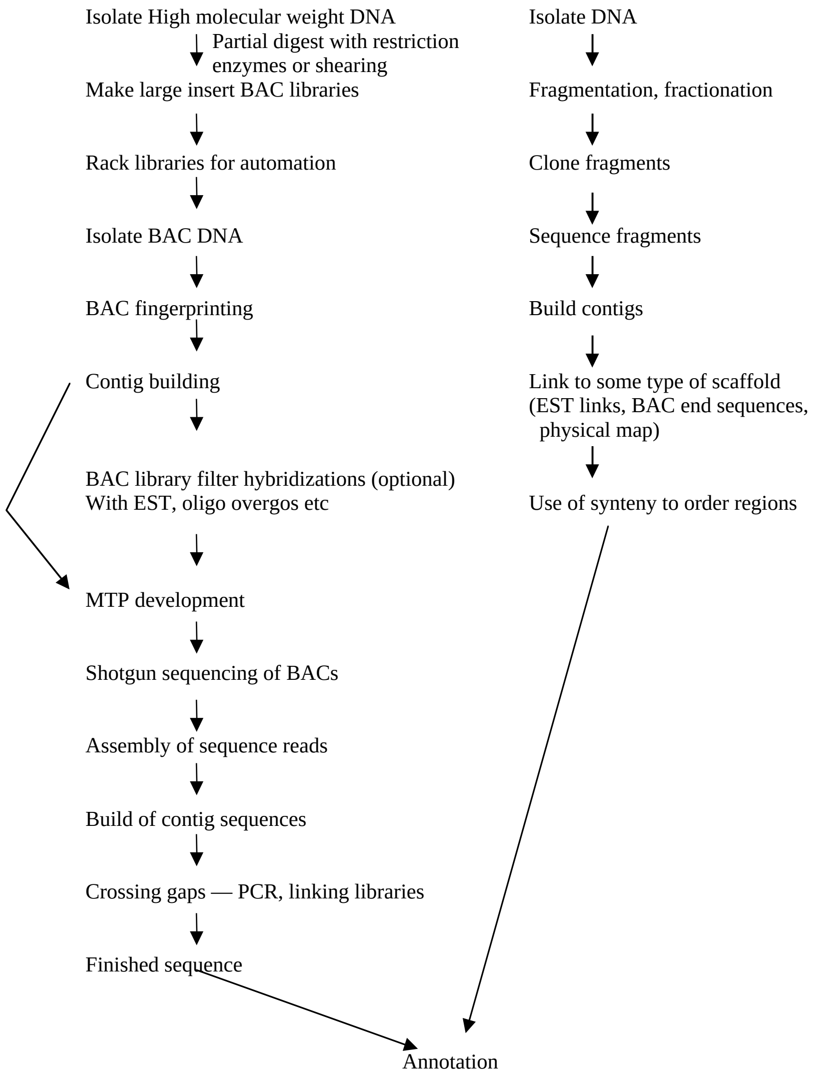


FIGURE 3.1. Flowcharts for various sequencing approaches.

Physical and Genetic Maps

The genetic map is an ordering of the regions of the genome according to the rates of recombination that occur along each of the chromosomes, the genetic distance being the frequency with which recombination occurs between two points on the chromosome. There is not a specific relationship between the genetic distance and the physical distance between the same

two points because the recombination rate is not constant over the length of the chromosomes. A more extreme example of a lack of correspondence between genetic and physical distance is found in humans, where the genetic maps for males and females have the same linear order but do not have the same values for the distances between equivalent points because there is a difference in recombination rate in the two sexes, the rate being higher in females than in males. The density of markers on the genetic map for most organisms has increased rapidly with the use of DNA-based polymorphisms. The genetic map is generated by the scoring of numerous polymorphisms in populations segregating for those markers. The larger the number of individuals scored in the population, the more precise the location of any particular marker becomes and the greater confidence can be placed on the specific order of a series of closely linked markers.

The physical map, on the other hand, is the linear order of the sequences linking the ends of each chromosome with overlapping regions of DNA. This is done by fragmenting the whole genome into pieces and then trying to reassemble the pieces into a single continuous molecule for each chromosome. When complete, each chromosome would consist of a single composite strand made up from a large number of smaller, overlapping fragments. In generation of the physical map, some of the regions will provide special problems. For example, regions that contain long sequences of tandem repeats, such as the genes for the large ribosomal RNAs, cannot be linked in a linear order, because the array is longer than any single subfragment and each repeat is essentially identical. However, a knowledge of the number of repeats and the number of chromosomal sites over which these repeats are spread can facilitate an estimation of the length of the region containing the repeats.

Generating a Physical Map

The current technology for generating a physical map usually involves ordering a series of BAC clones (Marra et al., 1997). The BAC libraries can be generated by using a number of different restriction enzymes (see Chapter 2) or by random shearing. The clones from these libraries are then fingerprinted. The fingerprinting involves isolating the BAC DNAs, digesting them with a restriction enzyme, and running the fragments generated on a gel. The overlap between different BACs is calculated from the number of identically sized fragments that they have in common (Soderlund et al., 2000). When a sufficient overlap between two BACs is found (the cutoff is user defined), those two BACs are assumed to have the region containing the identically sized bands in common and are therefore placed in a contig (contiguous sequence) (Figure 3.2). The comparisons and assembly can be automated with the use of image software to analyze the gels, and the fin-

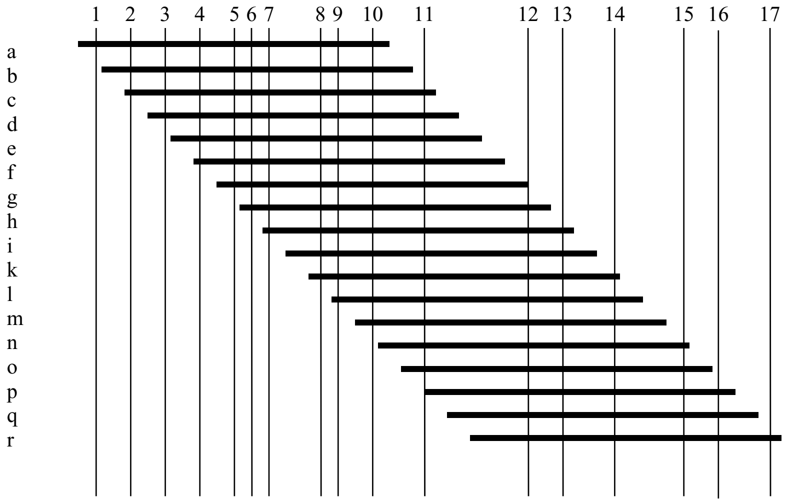


FIGURE 3.2. Illustration of the DNA fingerprinting method. Horizontal lines (*a–r*) represent individual BAC clones that have been aligned based on shared restriction fragment patterns. Vertical lines (*1–17*) represent the positions of restriction sites that were used. For actual data, the comparison usually includes 20–40 bands. In the figure the overlap between clones *a* and *b* is very likely to be correct. However, without the intermediate clones *b* through *h*, the overlap between *a* and *i* would be tenuous at best. As can be seen from the diagram, to be certain about the relative overlaps the same region must be sampled multiple times, with 13 different BACs contained within the length of *a*.

gerprint contigs (FPC) software is used to assemble the contigs (Soderlund et al., 2000). All the data generated from thousands of BACs are assembled, and the assembly constitutes the physical map. This assembly must be checked by confirming that the genes, markers, or sequences that are known to be on the BACs assembled into specific contigs are actually present in close proximity on the genetic map. Ultimately, all of the data that have been accumulated from the molecular mapping exercises can also be placed on the physical map. Thus where a unique molecular marker is hybridized to the BAC library, the selected fingerprinted BACs to which it hybridizes are immediately anchored to the position of the chromosome assigned to the marker. If the genetic and physical maps disagree, then the conflict must be resolved to determine which assembly, the genetic map or the BAC contig, is correct. Such physical maps based on fingerprinted contigs have been constructed for a number of species including humans, *Arabidopsis*, and rice.

EST markers that have been mapped to BAC clones can be entered into the FPC database. These data will help the assembly process, as well as placing the ESTs on both anchored and unanchored contigs (Soderlund et

al., 2000). The FPC software can use both the fingerprints and the markers in generating the assembly, so if two clones share a marker, then a less stringent overlap based on fingerprints will still be recorded as an overlap. Because FPC can generate incremental updates, the contigs can be built as the data are generated, rather than having to await all the data and having a massive one-time final build of the physical map.

How many BACs are needed to achieve such a physical map (assuming that there are no structural impediments to achieving the overlapping sets such as identical duplicated regions)? If we assume that the BAC library was generated with the average insert size of 125 kb, then a complete genome would be contained in about 1000 BACs for *Arabidopsis* (a genome size of 125 Mb), 3512 BACs for rice (genome size of 439 Mb), 21,728 BACs for maize (genome size of 2,716 Mb), and 128,000 BACs for wheat (genome size of 16,000 Mb). As can be seen from Figure 3.2 multiple sets of the genome must be fingerprinted, perhaps up to 20 times the number required for a complete genome so that enough representatives from each region can be sampled. Thus a twentyfold (20^x) oversampling would mean that the numbers for *Arabidopsis* rise to 20,000, for rice to 70,000, for maize to 435,000, and for wheat to 2,560,000 BACs. Therefore, as the genome size increases the number of BACs that must be fingerprinted to get some meaningful assembly of the genome also rises. Even with this level of oversampling, most of the plant genomes would not be assembled into the number of contigs that is the same as the number of chromosomes (the ideal result). In general, the number of contigs will be much larger than the chromosome number and the average size of the contigs much smaller than the length of the chromosome.

These contigs can then be placed on the genetic map by using molecular markers that hybridize to the BACs within a contig to determine the order of the contigs along the chromosome. Obviously, for this ordering along the genetic map to be successful, the spacing of the molecular markers has to be less than the size of the contigs so that at least one marker is present on each of the contigs. Alternatively, the BACs can be directly mapped onto the chromosomes with fluorescent in situ hybridization (FISH).

Fractionating the Genome

At present, the most important genomic information to acquire is the composition and number of the actual genes in a particular plant. The remainder of the genome does not hold quite the same importance because its contribution to the final phenotype is not thought to be substantial, and certainly less than that of the actual genes. Therefore, rather than trying to generate a complete genome sequence, strategies to enrich the regions that contain genes have been devised. These strategies usually rely on a demonstrated difference between the gene space, that is, the regions that contain

genes, and the rest of the genome. For plants that have very large genomes, this gene space is probably arranged in islands of gene-rich sequence separated by stretches of the genome that contain few, but greater than zero, genes (Panstruga et al., 1998). If such islands really exist and can be identified, then a large fraction of the genes could be isolated and sequenced apart from the rest of the "uninteresting" sequences. Characteristics that differentiate the genes from the rest of the nuclear DNA include the degree of methylation (Bird, 1986, 1992; Gruenbaum et al., 1981; Martienssen, 1999), the degree of repetition of the sequence within the genome, whether or not it is transcribed or contains an open reading frame, and, for maize at least, whether or not it is a target for transposable elements.

Methods of Fractionating the Genome

Expressed Sequence Tags (ESTs)

The last few years have seen an enormous growth of the number of ESTs in the databases for some of the major crop plants (Figure 3.3). This is clearly one source of genomic sequence for the genes. Obviously, any genes whose expression is either very low or restricted to tissues that were not sampled in the generation of the ESTS will be missed. Examples of this underachievement are the EST collections for human (3,500,000), *C. elegans* (150,000), and *Arabidopsis* (135,000) where only 35–65% of the genes predicted by genome sequencing were found in the EST collections. Additionally, various members of multigene families that do not differ in the region sequenced will be missed. Many of the ESTs have been placed on the genetic maps, but many have not because of the lack of any polymorphisms within the studied germplasm. For some plants, such as wheat, chromosome deletion lines that have already been developed can be used to localize these nonpolymorphic ESTs to a region of the chromosome. The development of maize/oat addition lines and radiation hybrids of these lines may serve the same purpose in corn (Kynast et al., 2001). However, this type of genetic resource is not likely to be developed for many other plant species.

Reassociation Kinetics

Up until the 1980s many genomes could only be characterized by reassociation kinetics, but this type of analysis went out of fashion with the arrival of easier, quicker modern molecular methods. These experiments physically separated the various classes of sequences, on the basis of the frequency with which they were present in the genome, by separating single- and double-stranded molecules after various incubation times. The more frequently a sequence was present in the genome, the more rapidly it reformed a duplex.

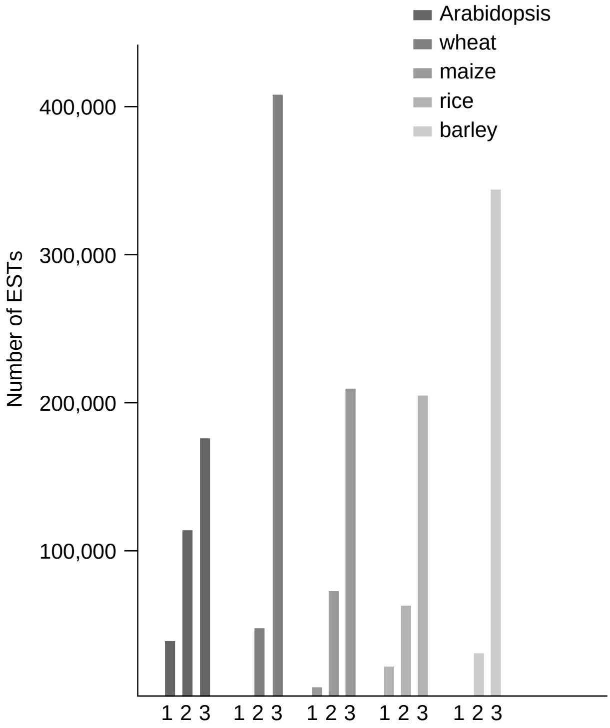


FIGURE 3.3. Increase in ESTs in dbEST. The different colors are for various species. 1—number of ESTs in 1998; 2—number of ESTs in 2000; 3—number of ESTs in 2003.

Therefore, a parameter designated Cot (for concentration times time) could be defined whereby various classes of repetitive sequences could be eliminated, or isolated physically, from the reaction. A reassociation experiment is carried out by shearing nuclear DNA into small fragments (200–500 bp) by high-speed blending and checking the fragment size by gel electrophoresis. The sheared DNA is precipitated, redissolved in the appropriate buffer, denatured, and allowed to reanneal at the appropriate temperature for various lengths of time. The single- and double-stranded fractions are physically separated with a hydroxyapatite column, and the amounts of the total starting DNA in each fraction are determined. The single-stranded fraction can be incubated again and the newly reassociated strands again isolated. This results in the physical isolation of the part of the genome that has sequences present with a particular range of copy numbers. The resulting Cot curve, such as that shown in Figure 3.4, can be used to assess the appropriate parameters to isolate a particular fraction of the genome. An example would be that by choosing the appropriate annealing

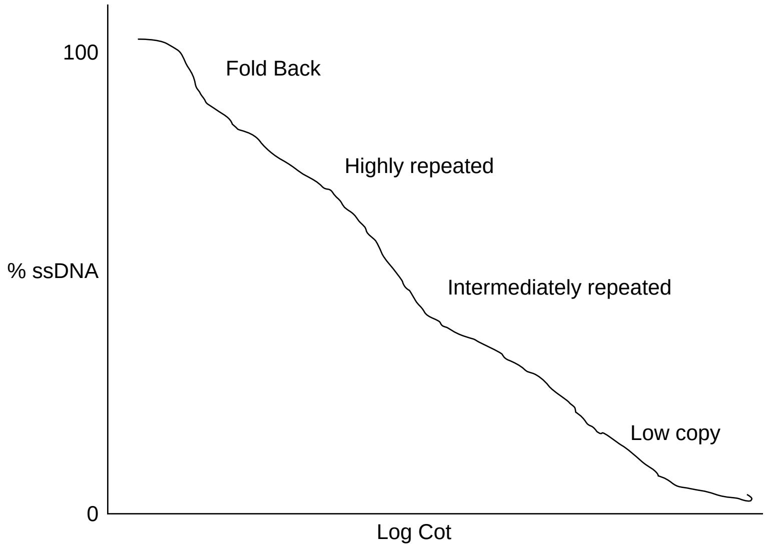


FIGURE 3.4. Cot curve for DNA from a higher plant of large genome size.

times and concentrations for the first and second incubations, those sequences that are present in the genome at between 1 and 10 copies could be isolated. Most of the most of the low-copy-number sequences are expected to be genes, but the proportion of all the genes that are in this fraction still must be determined. However, because of the interspersion of repetitive sequences adjacent to low-copy sequences, the fragments involved in the reassociation studies must be fairly short (500 bp or less for most complex genomes). A second consideration concerning the proportion of genes that will be in any particular high Cot fraction (the higher the Cot value, the lower the copy number of the sequences in the genome) is the stringency at which the reassociation is performed. In general, to achieve a reasonable rate of reassociation, the stringency of the reaction is set at about T_m – 25°C, which will allow about 25% of the nucleotides in any duplex to be mismatched. Therefore, even relatively distantly related sequences will appear to be present in multiple copies because of the cross-reaction under these conditions and so might be missing from the high-Cot fractions.

This is another example of a technique that had fallen into disuse but can be applied in a new context to provide vital information. Cot fractionation is a strategy for the fractionation of the genome that should be relatively unbiased and may result in the identification of genes not uncovered in any other fashion short of a whole genome sequencing effort (Peterson et al., 2002; Yuan et al., 2002).

Methyl Filtration

As with most higher eukaryotes, a portion of the cytosine residues at CpG or CpNpG sites in plant genomes are methylated (Bird, 1986, 1992; Gruenbaum et al., 1981; Martienssen, 1999). Methylation at these sites is known to modify DNA structure and regulate gene expression. This methylation is therefore variable within the genome, being lower in transcribed regions than in transcriptionally inactive regions. High rates of methylation (hypermethylation) are associated with transcriptionally inactive heterochromatin, whereas hypomethylation is usually associated with the transcriptionally active euchromatin. Therefore, elimination of the highly methylated regions would enrich the remaining sequences with genes. This discrimination becomes even more useful as the genome size increases. For example, most of the differences between methyl-C levels in corn and *Arabidopsis* can be accounted for by the relative amounts of methylated repeated DNA. Thus the level of gene methylation may be about the same in the two species. However, using the methylation status will render some parts of the genome that are of interest inaccessible. Thus, for example, the methylation status of genes, or portions of genes, can differ markedly between tissues and/or developmental stages. Furthermore, the association between methylation and expression is certainly not absolute. Therefore, the choice of tissue from which the library is constructed is important (Yuan et al., 2002).

The removal of methylated regions can be achieved at two different points in the cloning procedure. The first approach is with the use of a methylation-sensitive restriction enzyme, that is, one where the enzyme does not cleave the DNA if a cytosine in its recognition site is methylated. This approach has been used successfully to isolate low-copy-number sequences for molecular mapping probes, especially when using the enzyme *Pst*I. The second method takes advantage of the *E. coli* pathway that digests incoming sequences if they are methylated. Mammalian and plant DNA can be cloned more efficiently in bacterial strains that are mutant at the mcrA and mcrBC loci because these strains do not eliminate recombinants that contain methylated cytosines. Most of the time the cloning strategy is to get a high representation of all the sequences present. By using *E. coli* that still have a functional mcr system, all those recombinants containing methylated DNA will be eliminated, leaving a library highly enriched in nonmethylated sequences. Therefore, with the known distribution of much of the methylation in higher plants, the mcr system can be used to filter out methylated regions, resulting in an enrichment of genes at the expense of repeated sequences in such libraries.

Transposon Tags

For some well-characterized genomes such as that of maize, certain families of transposable elements are known to have a preference to transpose into

genes. Therefore, any method of isolating the regions adjacent to these elements will also isolate sequences enriched in genes. However, it is unlikely that this could be extended to use as a general method for sequencing a gene set for most higher plants because the distribution and characterization of the transposon content has not been done.

Selecting BAC Contigs Enriched for Expressed Genes

Rather than sequencing all the BACs that have been identified and placed on the physical map, those BAC contigs enriched for genes can be identified by using hybridization to overgo oligonucleotide probes (Han et al., 2000). Overgo probes are paired 24-mer oligonucleotides that contain an 8-bp complementary overlap. By annealing the oligonucleotides to each other and performing a [^{32}P]dCTP fill-in reaction, a labeled 40-mer is created that can be used for highly specific filter hybridization. This method has proven very efficient to construct sequence-ready BAC contigs for the human genome-sequencing project, where up to 40 distinct overgos have been used for simultaneous hybridization. The probes are designed from nonredundant cDNA or genomic sequences, but they still must be masked to eliminate repetitive elements. Therefore, some knowledge of the major families of repetitive elements for the species is necessary. The overgo probes are then designed from the masked sequence according to melting temperature, GC content, and potential for internal base pairing. The overgo probes are then hybridized to the BACs, and the labeled BACs are placed in the sequencing pipeline. Just sequencing the BACs containing expressed sequences will reduce the absolute amount of sequencing necessary. A possible second benefit to this strategy would arise if there were gene-rich islands, because the BACs selected in this way would have a high density of genes on them.

Strategies for Genome Sequencing

The gene enrichment or whole genome sequence approaches are not mutually exclusive. The most efficient ways of getting useful sequence from any genome of interest will probably include some combination of the enrichment, shotgun, and anchored sequencing methods. The following illustrates how such strategies can be incorporated into the two main sequencing approaches. What is important to realize is that the physical map is essential for genome analysis irrespective of whether minimum tiling path, whole shotgun, or gene enrichment strategies are being used.

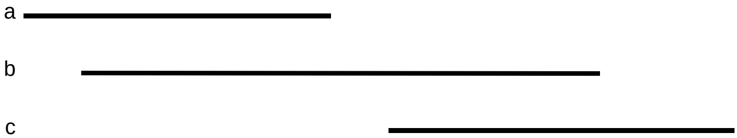


FIGURE 3.5. BAC contig where all the BACs have to be sequenced.

THE MINIMUM TILING PATH (MTP) SEQUENCING

This approach makes use of the physical map that must be developed before high-throughput sequencing is started. The physical map of overlapping BAC clones described above is the starting point. Obviously, it is not essential to get a complete physical map before beginning, but the regions that are the starting points must be defined. All the BACs in the same bin (for example, all the BACs in Figure 3.2, *a* through *r*, are in the same bin) can be readily identified. This information is then used to select a subset of BAC clones that will be used to determine the sequence of the contig. For example in Figure 3.2, the selection of *clones a*, *m*, and *r* would cover the whole of the contig with the minimum overlap. Here there will be substantial overlap between *m* and *r*, which will result in duplicate sequencing, but very little redundancy between *a* and *m*. However, with another contig as that shown in Figure 3.5, all three BACs would have to be sequenced, even though most of the sequence from BAC *b* would be redundant. As described below, other strategies for linking two BACs that are close together without the extra sequencing are available. Thus the minimum number of clones that are required to span a contig is determined by the confidence that can be placed in the alignment of the BACs in the contig. In Figure 3.2, the choice of the three BACs to sequence is relatively easy because of the large number of BACs with substantial overlaps that make up the contig. In Figure 3.5, *a* and *c* would be singleton BACs, that is, ones with no overlaps, if the BAC represented by *b* had not been fingerprinted. So even though they were in close physical proximity, they could not be placed relative to each other unless there was additional evidence. Such evidence could be the presence of molecular mapping markers on each of the BACs that were known to map very close together on the genetic map. However, as noted above, the correspondence between physical and genetic distances is not uniform, so the actual spacing of these two BACs could not be determined directly from these data.

BAC END SEQUENCING

All of the BACs that are fingerprinted can also be sequenced from both ends. This serves two purposes. First, it adds to the database of sequences for that organism. Current technology that can now generate up to a kilobase of sequence per run, so 2 kb of sequence is obtained from each BAC. Thus,

because each BAC is on the average 125 kb, this end sequencing would generate about 1.6% of the sequence of each BAC. Additionally, it is unlikely that a significant number of BACs would have overlapping ends, so all this sequence should be nonredundant, in the sense of representing various parts of the genome. However, if indeed the BACs are random regions of the genome, then the expectation is that only the same fraction of those end sequences would represent genes, because the genes only make up a small fraction of the whole genome. Thus the resulting sequence would represent about 30% of the total genomic sequence if the number of BACs included in the end sequence was a 20-fold coverage of the genome. A second purpose served is that the BAC end sequences can be used to generate the most efficient minimum tiling path. Starting from the sequence of a central clone in the BAC contig, the BAC end sequence data of all the other clones in the contig can be used to identify those with the minimal overlap. Thus the next contiguous clone to be sequenced is based on this minimal overlap. This approach is also called the sequence-tagged connectors (STC) approach. It would appear that the STC approach is the most sensitive in detecting the minimal overlaps and so involves the least amount of sequencing. One possible confounding factor in generating the minimal overlaps is where the minimal overlapping BAC clone ends in a repetitive sequence and so cannot be unambiguously identified. The next closest BAC that is uniquely defined then must be used. So if we go to Figure 3.6 then BAC 10 is chosen as the starting point. The whole of BAC 10 is sequenced and assembled. The BAC end sequences of the other clones in the contig are blasted against the sequence of BAC 10. BACs 7–12 all clearly overlap, and the end sequences are found in the sequence of 10. However, from the fingerprints it is not certain whether 6 and 13 also overlap with 10. If the BAC end sequences of 6 and 13 are found in the sequence of 10, then these two will be taken to continue the sequence of the contig. The caveat to this is that the end of 10 or the overlapping sequences in 6 and 13 must not be repetitive elements. Finally, all of the sequences can be blasted against the databases and any known gene homologies or sequenced marker homologies can be detected and used to anchor the particular BAC.

This MTP approach may be even more efficient at getting all the genes than expected because there is evidence that BAC libraries generated with restriction fragments are not random. An inspection of restriction enzyme digests of genomic DNA from most higher plants, with the commonly used six-base recognition site enzymes for BAC library generation, does not give the expected distribution of fragments around a size of 4096 bp. In most cases there are large regions without any sites, and these are overrepresentative of the heterochromatic fraction. Therefore, libraries made with restriction enzymes will be somewhat enriched for low-copy-number (genic) sequences. The use of random-sheared BAC libraries would be one way to overcome gaps containing much of the heterochromatic regions that may be

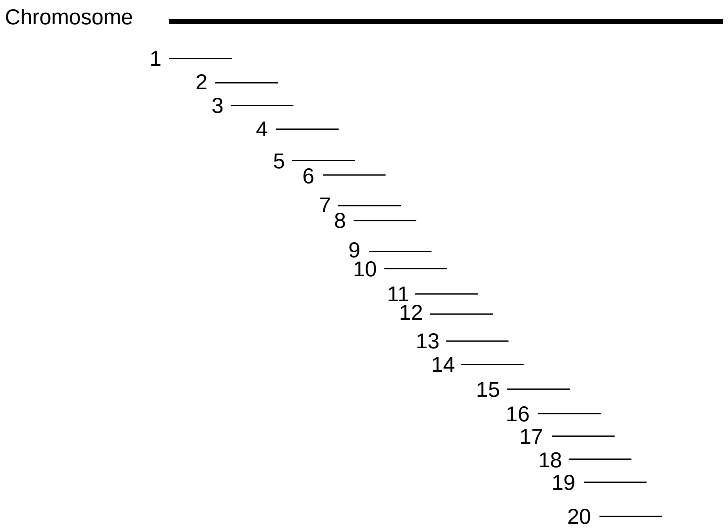


FIGURE 3.6. Use of BAC end sequences to determine minimum tiling path.

lost during the production of BAC libraries from restriction digest of DNA. However, if the heterochromatic regions are of less interest, then the restriction enzyme-digested DNA should be used for the BAC library construction.

GENOME FRACTIONATION APPROACHES

Here a physical map is not a prerequisite for collecting the sequence information. The initial activity in every case is to generate libraries from some form of a fractionated genome.

METHYL FILTRATION LIBRARIES

Shotgun libraries from crop plants and their propagation in strains of *E. coli* that restrict methylated DNA, thereby greatly enriching these libraries for gene-containing regions, have been developed by Orion Genomics (Rabinowicz et al., 1999; Yuan et al., 2002). The library construction must reduce or eliminate the organellar genomes that are essentially unmethylated. From normal green tissue, up to 5% of the total DNA could be derived from the chloroplast. If the gene space that is desired was only 15% of the total genome, then the organellar "contamination" would be about 25% of all the clones. Therefore, starting from isolated nuclei from nongreen tissue should reduce the unwanted chloroplast "contamination." Although

the expressed sequences are hypomethylated, they are not necessarily completely unmethylated. Therefore, the DNA that is cloned must be short enough so that no methylated cytosines are present in the clones of interest and must also be random. This can be achieved by shearing to low molecular weight (<500 bp), with the fragments then size selected, ligated into the appropriate vector, and transformed into the appropriate *E. coli* host strain. The resulting clones can be picked into the most appropriate format, for example, into 384-well plates for automated plasmid isolation and sequencing.

The initial evidence indicates that the filtration may be more successful when used with monocotyledonous than with dicotyledonous plants. Thus in rice, 128 mb remained after methyl filtration, giving an enrichment factor of nearly 4, whereas in wheat only 874 mb remained after filtration, giving a 19-fold equivalent enrichment. However a threefold enrichment was observed in canola, soybean, and tomato, the same value irrespective of the genome size. Thus it would appear that as the genome gets larger for monocots the filter power increases, whereas as the genome size gets larger for dicots the filter power is reduced (Bedell, 2002).

Methylation Restriction Libraries

Methylation-sensitive restriction enzymes such as *Pst*I have been used to produce libraries enriched in hypomethylated sequences. Any approach that uses methylation-sensitive restriction enzymes must take into account the size of the inserts that will be obtained and the completeness of the coverage of the subsequent libraries. However, in a genome with little previous genomic sequence, a shotgun sequencing project using *Pst*I clones may yield a large amount of data on genes of that organism. The resulting clones can be picked into the most appropriate format, for example, into 384-well plates for automated plasmid isolation and sequencing.

Cot Fractionation-Based Libraries

A preliminary Cot analysis can be performed if required to define the conditions to be used to isolate particular, interesting regions of the genome, although many species have already been characterized in this manner. From the Cot analysis the genome size can be estimated as well as the number and relative proportion of the repetitive and single-/low-copy components. The reassociated fraction selected is already double stranded, although the fragments may not have flush ends suitable for cloning. Therefore, these need to be blunted and then cloned in standard vectors. The resulting clones can be picked into the most appropriate format, for example, into 384-well plates for automated plasmid isolation and sequencing.

Sequencing and Data Processing

The sequencing activity is essentially similar for all of the approaches. With the MTP, the BACs are usually shotgun cloned into a new vector and the resulting fragments sequenced to a coverage to ensure that every possible fragment is included. The sequences of all the fragments are then assembled into the linear order in which they were in the original BAC clone. Obviously, if a BAC clone contains copies of a repetitive sequence, then the assembly of the complete BAC sequence will be more difficult.

Most of the high-throughput sequencing centers are based on the use of ABI DNA sequencers and fluorescent DNA sequencing chemistry with software for base calling, trace trimming, and quality assessment to ensure a uniform data standard for genome assembly (Ewing and Green, 1998; Ewing et al., 1998).

The first step in generating an assembly of shotgun sequence data is to group the shotgun sequences together into clusters of overlapping sequences. The second step is usually to check the quality of the sequence reads and then to identify possible contaminating vector or other sequences missed when the initial trimming of the sequences was done. Cloned sequences are then usually compared with sequences in public and other accessible databases, such as the GenBank Nr (nonredundant) and EST Databases and classified according to the nature and significance of their BLAST hits. Any repetitive sequences that are already known for the species under consideration can be masked to eliminate these from the analysis.

These analyses depend on robust and reliable clustering protocols that are sufficiently stringent to avoid errors in the clustering of gene families but also relaxed enough for appropriate groupings to be found. All the sequences together with their assembly and analysis can then be stored in an appropriate database. Because whole genome sequencing is usually a cooperative effort distributed among many laboratories worldwide, a common and integrated information environment is essential so that detailed tracking and control of the information processes can be achieved.

The Most Efficient Approach(es)

Two competing strategies for the complete sequencing of large genomes have been described, one in which physical maps are developed followed by the selection of a minimal tiling path of clones to sequence, and the other using a whole genome shotgun (WGS) approach. A test of the power of the two methods was essentially carried out during the sequencing of the human genome. The International Sequencing Consortium used the human FPC map that had been developed by the International Mapping Consortium, and the draft sequence published by Celera used the WGS approach.

However, the Celera sequence included draft sequence from the public consortium and, as has been mentioned above, might not have been assembled as well without the public, anchored data. Because these two methodologies are not mutually exclusive, a combination of them would appear to be the best approach for the future sequencing of large genomes. The BAC-by-BAC approach results in easier to assemble sequences where ambiguities can be resolved and the location of the resulting sequence is known. The advantage of the WGS approach is that it is more amenable to high-throughput automation and also covers regions that cannot be cloned by BACs. The WGS approach, especially if the aim is not to produce finished sequence, will be much less expensive.

Likely Targets for "Complete" Genome Sequencing

Sequencing strategies must be developed to account for the information described in Chapter 2 related to the structure of the genome of the particular plant under investigation. For example, in small genomes where much of the genome is present in long stretches of genes with relatively few repetitive sequences, such as *Arabidopsis thaliana*, the acquisition and analysis of the sequence data will be less complicated than in a large genome with a very high proportion of complex repeats and many related copes of a specific gene. Shotgun sequencing, the acquisition of random reads of sequence, could be assembled in the former case, whereas in the latter case it would be much more difficult. So the question is, what is the added value of such endeavors, or will much of the sequencing of additional genomes be a rediscovery or confirmation of rules gleaned from the *Arabidopsis* and rice sequence data? Despite the reduction in sequencing costs, generating enough sequence reads over large genomes is still an expensive proposition. This cost is then compounded by the problem of assembly, which is still a major concern is the cases of complex genomes and may be even more intractable for many of the very large polyploid plant genomes. With more than a single copy of a gene present in the genomes that are closely related, how can the different members be distinguished and differentiated from sequencing errors? If two copies of a gene are only minimally different, then how do you distinguish them? If the level of similarity is set too low, multiple copies will be merged into a single gene, whereas if it is too high, then sequencing errors will generate additional phantom copies. These are some of the considerations that come into play when trying to deal with the sequencing of complex plant genomes. Physical mapping followed by sequencing of the overlapping BAC clones was the strategy adopted for *Arabidopsis thaliana*. The sequence was then assembled into the final map. Even in this relatively simple genome, there are still runs of repetitive sequence that were not fully sequenced, although the lengths of these are regions are

known. However, adopting this strategy requires the whole physical map and a detailed genetic map before the sequence can be assembled on the scaffold. With larger genomes it may not even be possible to develop the physical map, much less be able to obtain the whole genome sequence.

Relatively few plants have huge sequence databases associated with them, and even fewer have large tracts of contiguous sequence. Comparisons across species can be very valuable, but the degree of relatedness in the comparison affects the kinds of questions that can be asked. In general a rule of pairs has been developed to allow the characterization of processes that have evolved within lineages. A wide range of plants have been characterized genetically and physiologically, and so all are potential subjects for detailed extensive genome sequencing. Sequencing projects already under way include those for the cabbage *Brassica oleraceae*, for two legumes, *Lotus japonicus* and *Medicago truncatula*, as well as for maize. Discussions are also under way to develop sequencing projects for soybean, tomato, barley, and banana.

The maize research community has organized and developed a plan for the needs of that community. A genome sequence was their highest priority (Bennetzen et al., 2001). In a sense this was similar to the way in which both the *Arabidopsis* and rice sequencing projects were started. Maize has a wealth of genetic data collected over the last century, and many important agronomic traits have been mapped. Therefore, the information derived from a genome sequence would be applicable to crop improvement as well as a basic understanding of how plants work. The initial efforts for maize are to generate the sequence of the gene space, rather than a complete genome sequence, mainly because of financial considerations as to the cost of a complete maize sequence.

Medicago truncatula, a relative of alfalfa and also a legume, already has an international effort to obtain a whole genome sequence (http://medicago.toulouse.inra.fr/EU/documents/whitepapergensequ.pdf). This species has prominence as a model legume, so an understanding of all the genes should aid in the understanding of the control of the symbiotic relationship between legumes and *Rhizobia*. The species also has a relatively small genome that should make the task of assembly easier but still not trivial. The first rounds of shotgun sequencing in this species have resulted in the complete sequence of the chloroplast and the definition of repetitive sequence classes. A BAC-anchored effort is also under way.

The poplar genome is a subject of a shotgun sequence to be done by the Joint Genome Institute (http://genome.jgi-psf.org/poplar0/poplar0.home.html). This effort is to generate a large number of reads, but the poplar community will have to assemble the sequence as a separate effort.

The other species listed above are also likely only to have a gene space sequencing effort initially because the genome is very large or the research

community is insufficient to support a whole genome sequencing effort. Tomato is likely to become the reference species for the Solanaceae and barley one of the grass reference species.

How many genes do plants have? If the actual size of the gene space is relatively constant, then the need for gene enrichment rises dramatically as the genome size increases. Thus wheat may have less than 5% of its genome as genes, and sequencing the rest may not be particularly useful or instructive. However, it is still a major and expensive undertaking to develop the sequence resources for a particular species. As the databases are populated with sequence data and the understanding of how genes are organized and distributed, it may become possible to devise improved strategies for generating genomic sequences efficiently. Until then, for most species syntenic relationships, supplemented by some EST sequence and perhaps the sequences of a few selected BACs, will have to suffice. In many cases, sequence data for most species will be generated in response to specific questions concerning the structure of particular genes across the plant kingdom where the sequences are generated with PCR and primers in conserved regions of the genes under study.

However, despite all the hurdles, the amount of sequence from higher plants will continue to rise at an accelerating rate for the foreseeable future. The first steps down this path were supported by funds from the National Plant Genome Initiative (through the Plant Genome Research Program at the National Science Foundation and the USDA). The work done with the additional funding provided under this umbrella has fundamentally altered the way plant research can and will be done.

REFERENCES

Bennetzen, J. L., V. L. Chander, and P. Schnable (2001) National Science Foundation-Sponsored Workshop Report. Maize Genome Sequencing Project. *Plant Physiol.* **127**, 1572–1578.

Bird, A. P. (1992) The essentials of DNA methylation. *Cell* **70**, 5–8

Bird, A. P. (1986) CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209–213.

Butcher, J. (2001). "Celera's method failed", says Human Genome Project *Lancet* **357**, 531.

Ewing, B., L. Hillier, M. C. Wendl. and P. Green (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**,175–185.

Ewing, B. L., and P. Green (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194.

Goff, S. A., D. Ricke, T. H. Lan, G. Presting, R. L. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchinson, C. Martin, F. Katagiri, B. M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. P. Zhong, T. Miguel, U.

Paszkowski, S. P. Zhang, M. Colbert, W. L. Sun, L. L. Chen, B. Cooper, S. Park, T. C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. S. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R. M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant, and S. Briggs (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp japonica). *Science* **296**, 92–100

Green, P. (1997) Against a whole-genome shotgun. *Genome Res*. **7**, 410–417.

Gruenbaum, Y., T. Naveh-Many, H. Cedar, and A. Razin (1981) Sequence specificity of methylation in higher plant DNA. *Nature* **292**, 860–862.

Han, C. S., R. D. Sutherlan, P. B. Jewett, M. L. Campbell, L. J. Meincke, J. G. Tesmer, M. O. Mundt, J. J. Fawcett, U.-J. Kim, L. L. Deaven, and N. A. Doggett (2000) Construction of a BAC contig map of chromosome 16q by two-dimensional overgo hybridization. *Genome Res*. **10**, 714–721.

Kynast, R. G., O. Riera-Lizarazu, M. I. Vales, R. J. Okagaki, S. B. Maquieira, G. Chen, E. V. Ananiev, W. E. Odland, C. D. Russell, A. O. Stec, S. M. Livingston, H. A. Zaia, H. W. Rines, and R. L. Phillips (2001) A complete set of maize individual chromosome additions to the oat genome. *Plant Physiol*. **125**, 1216–1227.

Marra, M. A., T. A. Kucaba, N. L. Dietrich, E. D. Green, B. Bownstein, R .K. Wilson, K .M. McDonald, L. W. Hillier, J. D. McPherson, and R. W. Waterston (1997). High throughput fingerprint analysis of large-insert clones. *Genome Res*. **7**, 1072–1084.

Martienssen, R. A. (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the corn genome. *Nat. Genet*. **23**, 305–308.

Panstruga, R., R. Bueschges, P. Piffanelli, and P. Schulze-Lefert (1998) A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. *Nucleic Acids Res.* **26**, 1056–1062.

Peterson, D. G., S. R. Schulze, E. B. Sciara, S. A. Lee, J. E. Bowers, A. Nagel, N. Jiang, D. C. Tibbitts, S. R. Wessler, and A. H. Paterson (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery *Genome Res*. **12**, 795–807.

Rabinowicz, P. D., K. Schutz, N. Dedhia, C. Yordan, L. D. Parnell, W. R. McCombie, and R. A. Martienssen (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet*. **23**, 305–308.

Soderlund, C., S. Humphray, A. Dunham, and L. French (2000) Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res*. **10**, 1772–1787.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 795–815.

Venter, J. C., H. O. Smith, and L. Hood (1996). A new strategy for genome sequencing. *Nature* **381**, 364–366.

Yuan, Y. N., P. J. SanMiguel, and J. L. Bennetzen (2002) Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *Zea Mays*. *Genome Res*. **12**, 1345–1349.

CHAPTER

4

Gene Discovery

Overview

The ultimate aim of genomics is to identify the structure and function of all the genes of all organisms. To be able to identify an end point in the attempt to characterize all the genes in plants, it is important to first define what we mean by a gene. The molecular definition of a gene has changed over the years. Initially, a gene was defined as the nucleic acid sequence that defines a peptide. Now the definition has to encompass many more features. These features include the presence of gene families within a plant, alternative splicing, RNA functioning without ever being translated into a protein, and other confounding factors that together make a simple universal definition more difficult. How different would two members of a gene family have to be in their structure, their pattern of expression, and/or the substrate specificity of their product for them to be classified as two different genes? For example, plants make many secondary products, groups of which are related and made by enzymes that are very similar. Although, more than 1000 monoterpenes are made by minor variants of a single enzyme (The *Arabidopsis* Genome Initiative, 2000), the number made by different species can vary widely. Should each of these variants be counted as a different gene, or the whole group as a single gene? Can the various distinguishable members of a gene family expressed only in particular tissues or at specific developmental stages be considered as separate genes because their expression is separated in time and space? Without a redundancy of function, mutations in each one of the members of the family may result in different phenotypes. Therefore, each member of the gene family could be mapped as an independent genetic locus, and so on this definition they could be considered as individual genes (some resistance gene clusters contain related genes that

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

would be classified as members of a gene family on the basis of sequence similarity, but each of the genes has a different specificity and is considered as a gene in its own right). Added to these considerations are the posttranslational modifications that can occur on the proteins, which can result in modifying their function.

The identification of genes can start at various points in the process of gathering information about genomes. They can be identified from their presence in populations of RNAs, from an analysis of genomic sequence data with gene finding programs, from comparisons with the genomic sequence data from related organisms, or from their disruption with the subsequent appearance of a phenotypic variant.

Therefore, the studies that enable gene discovery include:

- The cloning and sequencing of cDNAs to generate expressed sequence tags
- The cloning and sequencing of full-length cDNAs for understanding gene structure
- The annotation of whole genomes
- The use of syntenic relationships between species
- The use of transposons or T-DNA for gene tagging and insertional mutagenesis
- The use of chemical mutagens
- The expression of abnormal RNAs that induce epigenetic suppression of homologous genes (gene silencing)

How can all the genes and their functions within a plant be discovered? An example of such an attempt is the *Arabidopsis* 2010 program supported by the National Science Foundation (Chory et al., 2000). The goal of this program is the determination of the function of all the genes (currently estimated at about 25,000) in *Arabidopsis thaliana* by the year 2010. This will be accomplished by using the continued annotation of the complete *Arabidopsis* genome sequence. A complementary effort using the rice genome sequence will enable an expansion and validation of the conclusions drawn from the *Arabidopsis* genome to encompass the whole higher plant kingdom. This explosion of DNA sequence information has created a unique opportunity to identify and investigate the function of all the plant genes.

Identification of Genes from Sequence Data

Expressed Sequence Tags

One measure of whether a sequence within the genome is a gene is its identification within an RNA population. Therefore, the cloning and sequencing

of RNAs is one way to identify genes. Short stretches of RNA sequences derived from cDNAs are referred to as expressed sequence tags (ESTs). Gene expression can be dependent both on the type of tissue and on the environment in which that tissue finds itself, so a wide sampling of many tissues in various growth and challenge conditions must be undertaken to identify all, or most of, the genes. As described in Chapter 2, the population of sequences represented in a cDNA library is a reflection of the abundance of the RNAs present in the tissue sampled. Therefore, genes that are expressed at low levels may be missed in projects that sequence cDNAs that are generated from unfractionated RNA.

The Generation of ESTs

A pipeline for the informatic analysis of EST sequences is shown in Figure 4.1 (adapted from http://www.zmdb.iastate.edu/zmdb/EST/assembly.html). The RNA is isolated and reverse transcribed into cDNA. The cDNA clones are sequenced by performing single-pass sequencing reactions from either the 5′ or 3′ ends of the cDNA or from both ends. The sequences are then clustered to identify a series of tentative unique genes (TUGs) or tentative contigs (TCs) that are present in the RNA population that is being sequenced. This clustering will identify the number of different RNAs present in the initial sample. The TUGs/TCs can then be compared with the current databases to identify which of these have already been described in the species under consideration and which are still absent from the current databases. Where hits to previously reported sequences occur, the new assembly is collapsed into a single consensus sequence and added to the database. Where hits occur to ESTs from other organisms, a possible function may be ascribable to the sequence.

The sequencing of any given sample is continued until the rate of finding new sequences drops below an acceptable level (for example 50% of all sequences are already present in either these data or from previous EST collections).

The cDNAs from various tissues or treatments are sequenced to determine the level of novel sequences that are in each sample. As before, sequencing continues until the rate of gene discovery drops below an acceptable level. This method will generate a huge redundancy of highly abundant RNAs. What are likely to be missed are those RNAs that are present in low abundances and those genes that are only expressed in specialized cells. Therefore, techniques facilitating the isolation of specific tissues or cells, such as laser capture microscopy and RNA amplification, may help in the identification of genes that are expressed at low levels or in very few cells. The dissection or isolation of specific tissues such as the peltate trichome glands (which are aggregates of 1–9 specialized cells suspended on a stalk above the aerial surfaces of many plants), where important secondary products are

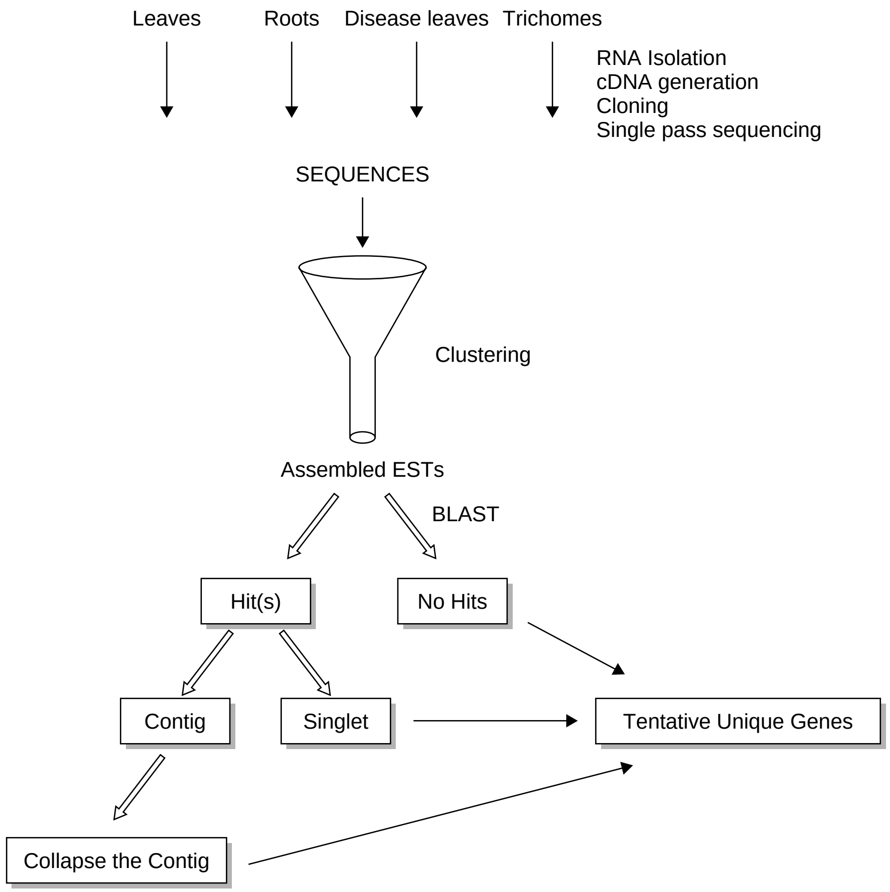


FIGURE 4.1. EST pipeline for the acquisition and assembly of sequences. Contigs are EST clusters with two or more member ESTs. Singlets are ESTs that are not significantly similar to any other ESTs. The combined contigs and singlets and those with no BLAST hits represent a set of unique EST clusters called tentative unique genes. They are labeled "tentative" to indicate that they are still subject to changes as new ESTs are added to the assembly.

synthesized, should lead to the isolation of the genes involved in these metabolic pathways (Wang et al., 2001).

The high-throughput EST sequencing approach represents a relatively low-cost method to identify a large number of transcripts in an organism as well as generating information about the patterns of gene expression specific to certain tissues, developmental stages, and physiological conditions. The value and importance of ESTs is indicated by the numbers in GenBank dbEST (http://www.ncbi.nlm.nih.gov/dbEST/dbEST'summary.

html). Release 2/14/2003 lists 14,411,241 ESTs, of which about 2,800,000 are from plants, with wheat, barley, and soybeans leading the list (Table 4.1).

As mentioned above an inherent problem of EST sequencing projects is the generation of redundant sequences. One way of reducing redundant sequencing is to enrich the RNA populations for low-abundance transcripts. A number of normalization and subtraction methodologies for enrichment of these low-abundance RNAs before cloning are described in Chapter 2. Alternatively, abundant cDNA clones can be removed before sequencing by screening high-density cDNA filters with labeled RNA. The clones that have strong hybridizations are eliminated, and the minimally-hybridizing clones are rearrayed and sequenced. There will always be some redundancy irrespective of the method of enrichment, but this can be managed informatically. The clustering and assembly of individual ESTs into TUGs/TCs will result in decreased sequence redundancy and a final consensus sequence that should be both more accurate and longer than any of the underlying individual ESTs in the database. The ultimate goal of EST projects is the development of a unigene set. The unigene set should eventually contain all the genes for the organism, but this complete compendium is unlikely to be assembled from just EST data because of the need to sample every possible tissue and find every transcript. However, the clustering algorithms will identify all the transcripts from a gene family and generate a consensus sequence from the EST data. Therefore, any information in the EST data that identifies the differing members of a gene family whose expression is restricted to a particular cell or tissue type will be lost in the development of the unigene set and have to be recaptured elsewhere.

The Generation of Full-Length cDNAs

Full-length cDNAs are an essential resource for the functional analysis of plant genes (Figure 4.2). The confirmation of the prediction of transcription units from genomic sequence data is dependent on validation with a full-length cDNA. Full-length cDNAs are also necessary for the correct identification of splice sites. The occurrence of alternate splicing events also must be confirmed by the identification of a full-length cDNA containing the alternatively spliced region. The full-length cDNA can be used in both homologous and heterologous expression systems to generate large amounts of protein for functional and structural studies to determine the function of the gene. In addition, sequencing of the full-length transcripts will allow the identification of RNAs from different members of gene families.

Full-length cDNA library construction is more technically challenging compared with EST generation. A full-length first-strand cDNA is not efficiently produced by reverse transcription, especially if the mRNA has a stable secondary structure. Libraries made from cDNAs, therefore, can

TABLE 4.1. EST numbers in dbEST as of 2/14/2003 for selected plant species

Wheat	415,589
Hordeum vulgare + subsp. *vulgare* (barley)	314,882
Glycine max (soybean)	308,564
Zea mays (maize)	197,921
Medicago truncatula (barrel medic)	180,939
Arabidopsis thaliana (thale cress)	178,464
Lycopersicon esculentum (tomato)	148,554
Oryza sativa (rice)	130,772
Solanum tuberosum (potato)	94,423
Sorghum bicolor (sorghum)	84,712
Lactuca sativa (lettuce)	68,188
Pinus taeda (loblolly pine)	60,226
Populus tremula × *Populus tremuloides*	56,013
Helianthus annuus (sunflower)	46,951
Gossypium arboreum	38,894
Lotus japonicus	33,124
Vitis vinifera (grape)	30,940
Ipomoea nil (morning glory)	25,899
Mesembryanthemum crystallinum (common ice plant)	25,446
Hordeum vulgare subsp. *spontaneum*	24,150
Populus balsamifera subsp. *trichocarpa*	23,717
Capsicum annuum (pepper)	22,433
Sorghum propinquum	21,387
Phaseolus coccineus	20,120
Beta vulgaris (beet)	19,617
Populus tremula	14,078
Gossypium hirsutum (upland cotton)	10,716
Prunus persica	10,185
Zinnia elegans	9,783
Triticum monococcum	9,572
Secale cereale (rye)	8,930
Lycopersicon pennellii	8,346
Oryza minuta	5,268
Brassica rapa subsp. *pekinensis* (Chinese cabbage)	4,316
Rosa hybrid cultivar	2,874
Prunus dulcis	2,858
Brassica napus (oilseed rape)	2,691
Citrus sinensis	2,623
Lycopersicon hirsutum	2,504
Mentha x piperita (peppermint)	1,316
Linum usitatissimum (flax)	1,299
Allium cepa (onion)	1,193
Cicer arietinum (chickpea)	23
Narcissus pseudonarcissus (daffodil)	1

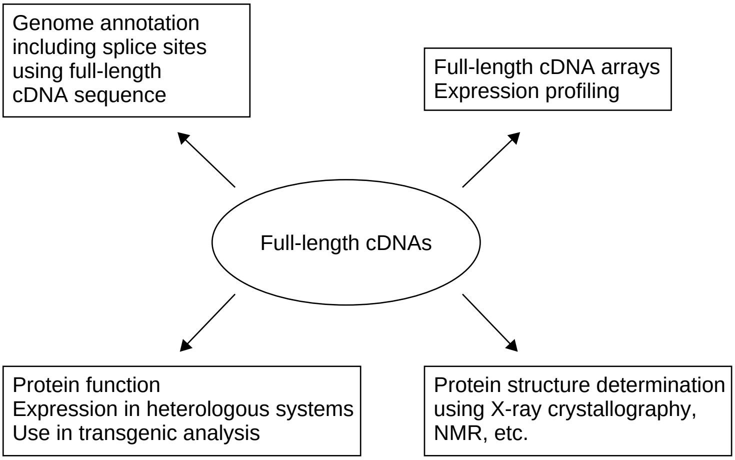


FIGURE 4.2. The contributions of full-length cDNA sequences to functional genomics.

contain both full-length and partial cDNAs. Efficient techniques exist for selecting only full-length cDNAs. One such method for constructing cDNA libraries with a high content of full-length clones involves starting from the first transcribed nucleotide (Figure 4.3). A biotin label for the cap structure, based on the principle that the cap site and 3′ end of mRNA are the only sites that carry the diol structure, has been developed. The diol groups at each end of the mRNA are biotinylated, and then the first-strand cDNA is synthesized. This synthesis is primed with a degenerate primer [XTTTTTTTT(Restriction site)]. The reaction mixture is then digested with RNase, and only the full-length cDNAs are protected from degradation of the unpaired mRNA. Therefore, the 5′ ends of all the partial cDNAs are removed (along with the biotin) as are the 3′ ends of all molecules. The full-length cDNAs are captured on streptavidin-coated magnetic beads, and the cDNA is released from the beads and the mRNA by treatment with RNase H and alkaline hydrolysis. The cDNA is then tailed with oligo(dG) that is used to prime the second-strand synthesis. Again, this primer also has an extension that includes a restriction enzyme site. After the second-strand synthesis the full-length cDNA is cloned with the restriction sites inserted with the first- and second-strand primers.

Full-Length cDNA Sequencing

The approach to obtaining the sequence of a full-length cDNA clone is different from that used to generate EST data. Many of the full-length cDNAs

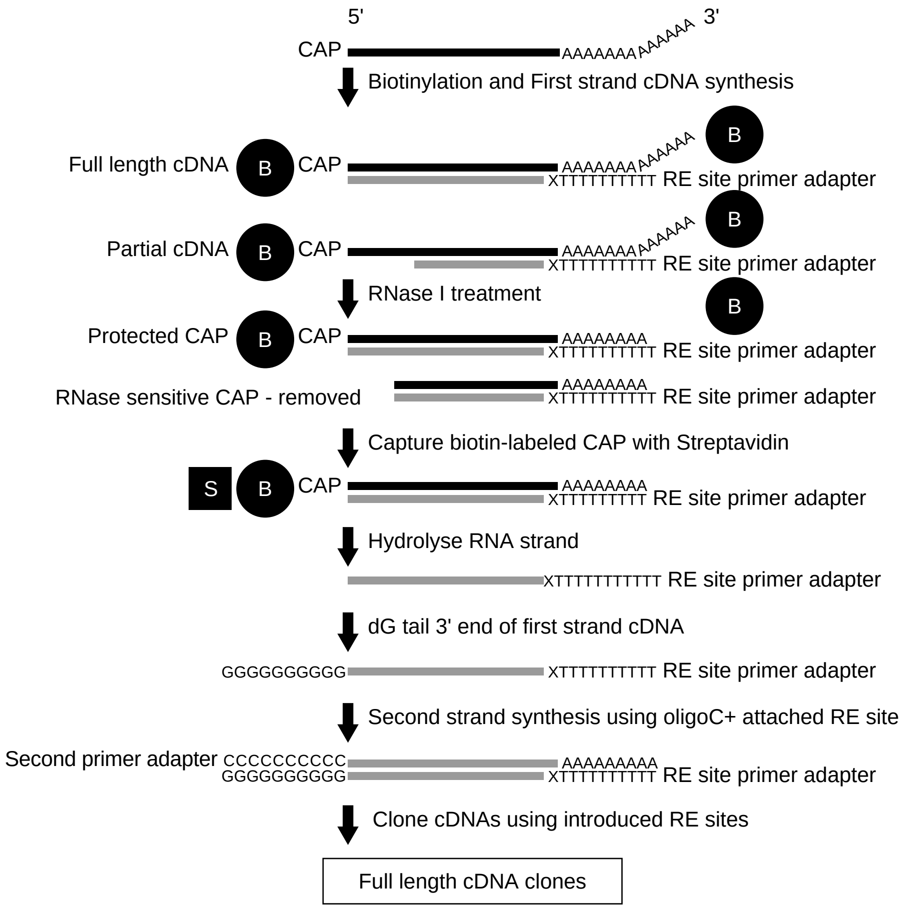


FIGURE 4.3. Schematic representation of a general strategy to select and clone full-length cDNAs: Biotinylation (B) of diol groups and synthesis of first-strand cDNA with a degenerate primer adapter including restriction enzyme (RE) site; RNase I digestion; full-length cDNA captured by streptavidin-coated magnetic beads (S); cDNA released from beads by RNA hydrolysis; first-strand cDNA tailed with oligo(dG); second-strand cDNA synthesis primed; cDNA cloned using the restriction sites introduced with the primers (Adapted from LabLink 2, No. 1 1998 http://www.cpg-biotech.com/lablink/lablink_2_1_1.html).

are likely to be longer than the reads resulting from sequencing both the 3′ and 5′ ends of the insert. Therefore, additional sequencing strategies are necessary to obtain the full-length cDNA sequence. The sequence must be of high quality to identify possible SNPs and the transcripts from different members of a gene family. The availability of such sequences is also important for the confirmation of genes identified through computational genome

annotation and for the identification of alternative splicing sites for RNAs. Three possible strategies for full-length sequencing are:

- Transposon mutagenesis: A transposon is randomly inserted, *in vitro*, into the cDNA insert, and primers designed from both sides of the transposon are used for sequencing. Sequencing a number of independent transposon sites will be sufficient to assemble the complete cDNA sequence.
- Concatenated cDNA sequencing (CCS): cDNA inserts are concatenated into a BAC-size molecule (100 kb) and then sheared, cloned, and shotgun sequenced by standard procedures. The full-length cDNA sequences are then assembled and edited with the appropriate software.
- Primer walking: Primers are designed from 5′ and 3′ end sequences and used for a second round of sequencing. Additional primers are then made and used until the whole contiguous cDNA sequence is obtained.

The availability of full-length cDNA clones opens new avenues for identifying gene function and protein interactions. As detailed more fully in Chapter 6, the use of tagged full-length cDNAs in transgenic plants can be a first step in isolating and identifying the protein complexes that exist *in vivo*. Transgenics can also be used to develop a protein atlas of where in the cell each of the genes is expressed. A full-length cDNA can be tagged with a dye and the tagged probe transformed back into the plant under the control of its native promoter. The site of the fluorescence will indicate the organ or tissue where the gene is expressed, as well as the cellular localization of the protein. The reintroduction of the full-length cDNA into a plant can also result in either overexpression or silencing of that gene. The subsequent phenotype that is observed provides clues as to the function of the gene. In addition, overexpression of such a gene, for which a full-length cDNA is available, can be accomplished in a heterologous system, such as yeast or *E. coli*, followed by *in vitro* studies of the protein function.

Thus, although ESTs are useful for the initial description of the transcriptome, the full-length cDNAs are a much more valuable resource. The availability of full-length cDNAs for many species will greatly enhance the rate of gene discovery, patterns of splicing, and the understanding of gene function in plants.

Genome Annotation

Among the first features to be placed on newly acquired assembled genomic sequences are the possible open reading frames and splice sites. These two

sets of data are combined to identify both already known and putative genes. An example of such an annotation is shown in Figure 4.4 (Ware et al., 2002). All the available information has been added to the sequence including the BAC end sequences, markers for the rice maps including the SSR markers, and positions of predicted genes, submitted genes, and other EST data from rice, maize, *Hordeum*, *Triticum*, and *Sorghum*.

In all gene predictions from genomic DNA the precise identity of the gene boundaries and exon-intron structure is hindered by the lack of supporting experimental evidence. Full-length cDNA sequences and bioinformatics software can produce insights on the structure of genes in chromosomal DNA. Therefore, full-length cDNA sequences are essential for confirmation of the predicted genes within a sequenced genome. Having a full-length cDNA enables the checking of both the extent of the coding region of the gene as well as the sequences immediately 5′ and 3′ from the coding sequence. In addition, having a full length cDNA makes it possible to train the gene finding programs so that the unknown regions of the genome can be more accurately annotated as far as the presence of genes is concerned. The availability of many full-length cDNAs and trained gene finding programs from a small number of model plants will also ease the identification of genes in partial genomic sequences of more exotic plant species.

Synteny

As the full genomes of *Arabidopsis* and rice are more precisely annotated, the finding and isolation of potential genes in other, less well-defined systems may be possible with reference to the position of the sequence in a particular cluster of genes. However, these predictions are likely to be complicated by the presence of multiple copies of genes, the divergence between paralogs and orthologs (see Chapter 1) in other species, and the micro- and macrorearrangements of the chromosomes over evolutionary time. Therefore, any candidates will need to be extensively characterized to demonstrate that they are performing the same function in both time and space.

Identification of Genes by Mutagenesis

The foregoing discussion has centered around identifying genes from the information gained from both the DNA and RNA sequences. Even if a DNA stretch is transcribed into RNA, it still must be shown that this RNA has a function, either directly as an RNA molecule or after the translation of the RNA into a protein. An alternative to the nucleic acid characterization is direct demonstration that this sequence has a function. This can be done by the reintroduction of the sequence into the appropriate plant or by trying to

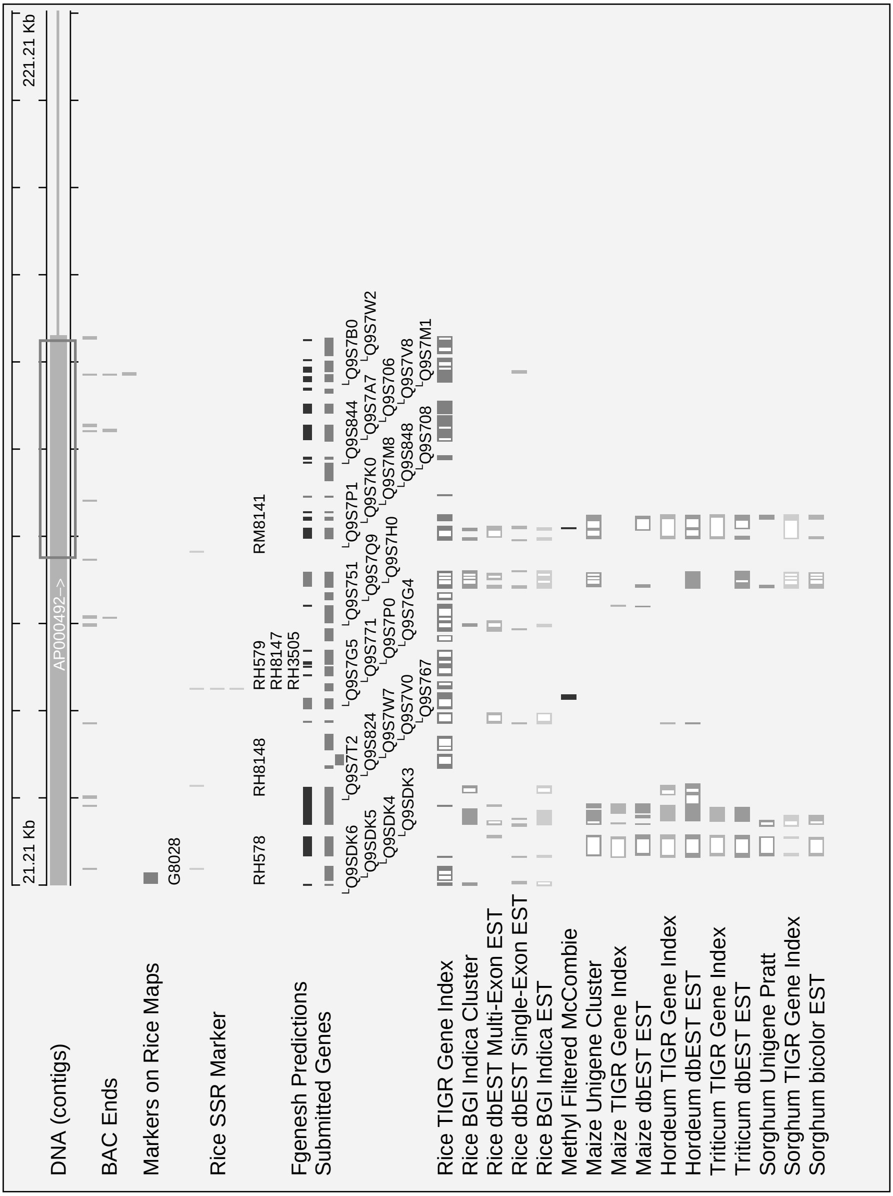


FIGURE 4.4. An example of an annotated DNA sequence from rice. Genome view of the BAC/PAC P0705D01 of rice from www.gramene.org.

knock out the gene by mutagenesis. The technology used to mutagenize genes and identify those mutants has been developed with both insertional mutagenesis (Azpiroz-Leehan and Feldmann, 1997) and the TILLING methodology (McCallum et al., 20002), which identifies single base changes in the gene of interest. A third method for disrupting gene function is by RNA interference (RNAi) silencing (Cogoni and Macino, 2000).

Insertional Mutagenesis

Gene Knockouts. Gene knockouts are where the activity of a gene has been eliminated. In plants the two major methods for generating these are by inserting either a T-DNA or a transposon sequence (Azpiroz-Leehan and Feldmann, 1997). Because of the lack of an efficient homologous recombination system in plants the technique of replacing a gene with a modified form is not currently available. Therefore, the elimination of activity by insertion is the most common method used to disrupt gene function. T-DNA insertion is the most generally applicable method because it can be used for any plant that can be transformed and regenerated. Because each transformant is an independent event with the T-DNA being relatively randomly inserted into the genome a large number of independent transformation events are needed to inactivate every gene. The need for the generation of large numbers of independent transformants therefore limits this technology to plant species or particular lines that are capable of being transformed in a high-throughput manner. In contrast, the advantage of using transposons is that they can then be activated and moved into many regions of the genome. Therefore, after the generation of a small number of lines with the transposon present, the transposons can be launched to move around the genome and generate insertions in every gene. This technique is most easily applied to maize because this is the plant from which most of the transposable elements have been isolated. The initial lines are, in essence, always available. The engineering of two-component transposon systems that include an inducible promoter will make this particular technique more widely applicable to a wide variety of other plant species.

Insertional Mutagenesis with T-DNA. The insertion of a T-DNA element into a chromosome can lead to many different outcomes:

- The insertion into the coding region can lead to partial or complete inactivation of the gene.
- The insertion into the promoter region can lead to any of following results:

- Complete inactivation of the gene
- Reduced expression of the gene
- Increased expression of the gene

A computer database has been established for *Arabidopsis* that contains the precise genomic locations of over 50,000 T-DNA insertions. Any gene of interest can quickly be found, if the collection contains a mutation in that gene, by performing a simple BLAST search. The database of these insertions can be found at http://signal.salk.edu/cgi-bin/tdnaexpress. If a T-DNA insertion is not found with this resource, the next resource to use would be the *Arabidopsis* Knockout Facility at the University of Wisconsin. The lines obtained from the *Arabidopsis* Biological Resource Center (ABRC) are used in the screening scheme outlined in Figure 4.5. A small number of other crop plants may have similar resources, especially rice, which is already well served with T-DNA insertion lines (Parinov and Sundaresan, 2000; Ramachandran and Sundaresan, 2001).

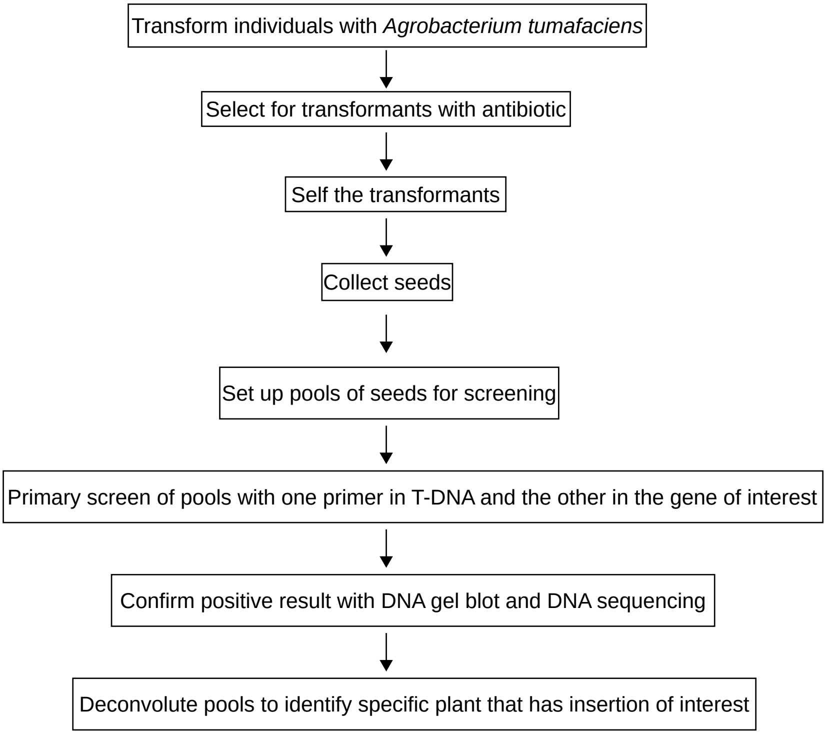


FIGURE 4.5. The flow of an experiment to generate T-DNA insertional mutants in genes of interest.

As well as using this insertion methodology to identify genes by knocking out their function, the use of either enhancer traps or gene traps can identify the promoters of genes as well as their patterns of expression. The process consists of placing a reporter gene in a vector whereby the reporter gene is only activated when inserted within a functional gene. The reporter gene has a visual phenotype, so the tissue specificity of the promoter region (and therefore the endogenous gene itself) can be identified directly. The reporter activation demonstrates the spatial and temporal expression of the disrupted gene. Because expression levels can be monitored in heterozygous plants, the gene trap system is useful for studying the patterns of most plant genes, including essential genes that cause lethal mutations when homozygous. Some examples are shown in Figure 4.6.

A finer dissection of various patterns within an organ has been demonstrated for enhancer trap GUS fusions in *Arabidopsis* roots (Figure 4.7).

Once again, because this approach requires many independent transformants to be made, it is limited to those lines in which high-throughput transformation is available. It has also to some extent been limited to plants that can be grown in large numbers to facilitate the screening of the transformants. Looking for a specific promoter that is expressed in a mature tissue in a maize plant is not going to be nearly as easy as looking for a similar promoter in *Arabidopsis*.

Transposon Mutagenesis. This method has been available in maize but is also being used in other organisms (Walbot, 2000). The scheme for such an experimental approach is shown in Figure 4.8. If the maize transposon Ac is used, the movement of the transposon is likely to be to relatively close sites on the same chromosome as the original insertion point. Therefore, a number of lines must be constructed with the Ac present at various chromosomal locations. The locations of the introduced Ac elements must be determined so that a series of starter lines can be selected with the Ac elements distributed around the genome. Then the appropriate starter line can be chosen that will have a high probability of generating an insertion in the gene of interest, that is, the Ac is near the gene of interest. As with T-DNA insertions, the construction of the transposon to function as an enhancer trap is also possible.

Targeting Induced Local Lesions In Genomes (TILLING)

Insertional mutagenesis is a fairly blunt tool with which to dissect gene function. It is often advantageous to have an allelic series that shows a gradation of gene activity. Point mutations in the gene can generate such a series, but these mutations are generally difficult to identify. A process called TILLING combines chemical mutagenesis with mutation screens of pooled PCR prod-

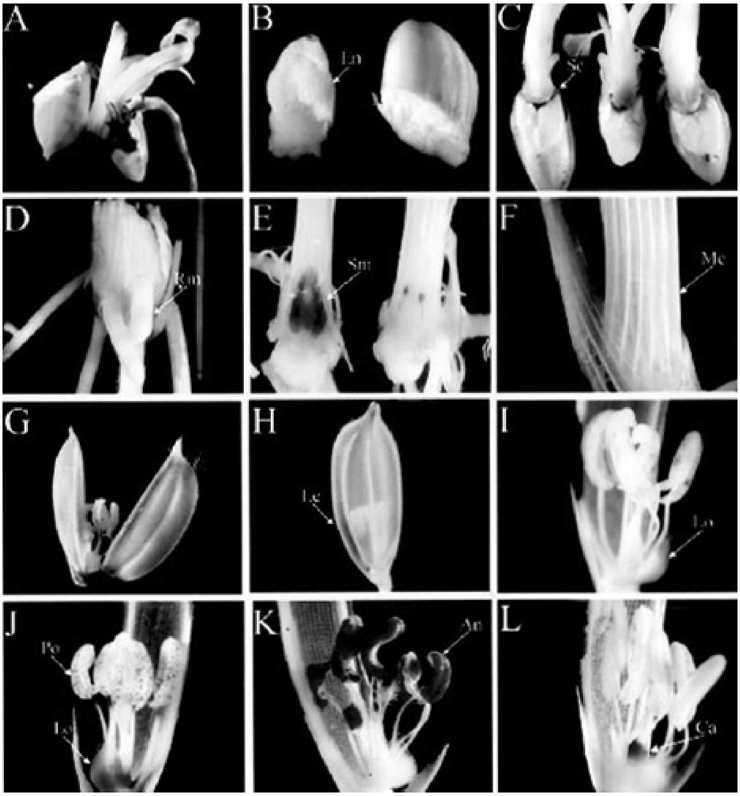
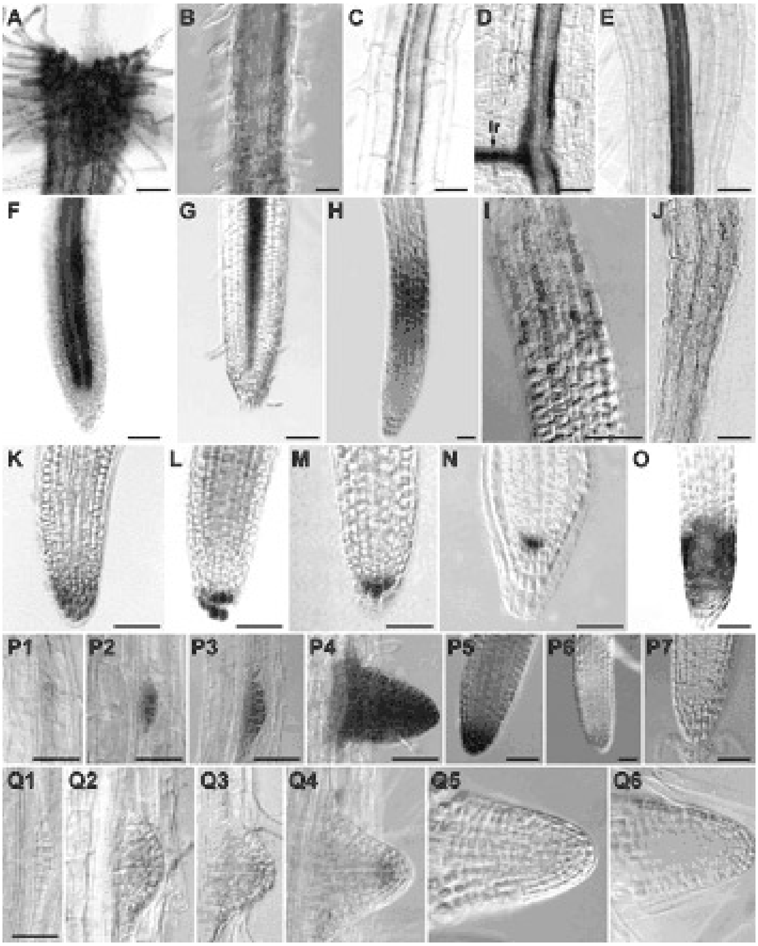


FIGURE 4.6. The expression patterns from activation tagging in rice to demonstrate tissue specificity. Analysis of GUS activity in transgenic rice plants. A, Line 1B-05504 exhibiting GUS activity in all seedling organs. B, Line 1A-10540 showing GUS activity in the endosperm. C, Line 1A-10620 showing GUS activity in scutellum. D, Line 1A-10919 with root meristem-specific GUS expression. E, Line 1A-10721 displaying preferential GUS activity in the shoot apical meristem. F, Line 1A-10601 exhibiting mesophyll cell-specific GUS staining in the sheath. G, Line 1A-24951 showing GUS activity in all floral organs. H, Line 1B-05625 with GUS activity in lemma. I, Line 1A-25114 with GUS activity in lodicules. J, Line 1B-24512 with GUS activity in pollen and lodicules. K, Line 1A-12905 displaying strong GUS activity in anthers. L, Line 1A-11513 exhibiting carpel-specific GUS expression. An, anthers; Ca, carpel; En, endosperm; Le, lemma; Lo, lodicules; Me, mesophyll; Po, pollen; Rm, root meristem; Sc, scutellum; Sm, shoot apical meristem. (Reprinted with permission from Jeong et al., 2002.)

ucts that allow a point mutation to be identified. The use of chemical mutagenesis can result in missense and nonsense mutant alleles of the targeted genes. Additionally, the use of chemical mutagenesis is applicable to any plant because it does not require transgenic or cell culture manipulations. In

the basic TILLING method (McCallum et al., 2000), seeds are mutagenized by treatment with ethylmethanesulfonate (EMS). The resulting M1 plants are self-fertilized, and DNA is prepared from the M2 individuals. To screen many individuals a pooling strategy is used. DNA samples are pooled, and pools are arrayed on microtiter plates and subjected to gene-specific PCR. High-throughput TILLING (Colbert et al., 2001) uses the CEL I mismatch cleavage enzyme (Oleykowski et al., 1998). The amplification products are incubated with an endonuclease that preferentially recognizes and cleaves mismatches in heteroduplexes that are formed in hybridizations between the wild-type and mutant alleles. The cleavage products are separated with a sequencing gel apparatus, and gel images are analyzed. If a mutation is detected in a pool, the individual DNA samples that went into the pool can



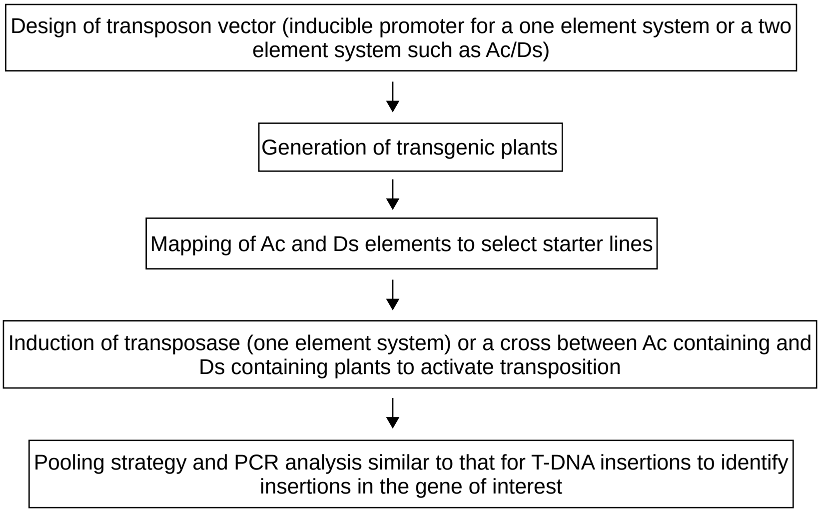


FIGURE 4.8. The generation of a population of transposon tagged plants for gene discovery.

FIGURE 4.7. Enhancer trap transposants with GUS activity in the root. A, UCR17, GUS activity in epidermis. B, UCR18, GUS activity in cortex. C, UCR19, GUS activity in endodermis. D, UCR8, GUS activity in phloem and pericycle. Arrow points to a lateral root (lr). E, UCR20, GUS activity in developing xylem. F, UCR21, GUS activity in outer layers of vascular cylinder and vascular initial cells in the RAM. G, UCR9, GUS activity in central layers of the vascular cylinder, originating in the vascular initials. H, UCR22, GUS activity in the zones of elongation and differentiation, localized to the epidermis and cortex. I and J, UCR23, GUS activity in trichoblast cell files in the elongation (I) and differentiation (J) zones. K, UCR24, GUS activity in the root cap. L through N, GUS activity in the bottom two tiers (L, UCR25), middle tier (M, UCR26), and initials (N, UCR27) of the columella root cap. O, UCR28, GUS activity in the lateral root cap. P and Q, Developmental series of lateral root initiation. P, UCR29, GUS activity in the first few cells of dividing pericycle (P1). GUS activity was visible throughout the lateral root primordia (P2–P5), was progressively restricted to the root tip (P6), and was not detected in mature roots (P7). Q, UCR30, GUS activity was not detected in early-stage primordia (Q1) but was visible before the lateral root primordia emerged from the primary root (Q2). GUS activity was restricted to the organizing RAM (Q3–Q4) and disappeared at later stages (Q5–Q6). Images were captured with differential interference contrast microscopy of root whole mounts. Scale bars = 50 μm. Scale bar in Q1 refers to Q1 through Q6. (Reprinted with permission from Geisler et al., 2002.)

be individually analyzed to identify the individual that carries the mutation. Once this individual has been identified, its phenotype can be determined. A potential problem with this method is that any one individual will carry multiple mutations. Genetic analysis is therefore necessary to confirm that any observed phenotypic alteration is associated with the mutation in the target gene and not with another mutation elsewhere in the genome.

RNAi (RNA Interference)

All gene disruption approaches have some inherent limitations. For example, it is difficult to identify the function of redundant genes or the functions of genes required in early embryogenesis or gametophyte development.

One way to overcome the redundant gene problem is to simultaneously inhibit all the members of a gene family (gene silencing). RNAi refers to the function of homologous double-stranded RNA (dsRNA) to specifically target a gene's product, resulting in null or hypomorphic phenotypes. As

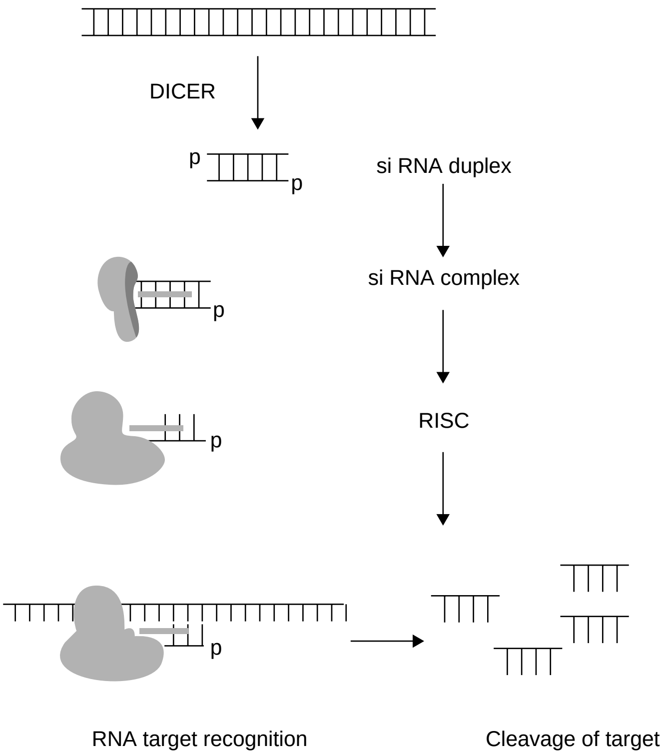


FIGURE 4.9. The mechanism of RNAi. (Figure provided by Sirna Therapeutics and Adapted from Genetic Engineering News vol 22, No. 21 page 3.)

long as the interference is targeted to a region of the gene that is conserved within all the members of the gene family, all members of the family will be similarly inhibited (Tang et al., 2003).

The most interesting aspects of RNAi are the following:

- dsRNA, rather than single-stranded antisense RNA, is the interfering agent.
- It is highly specific.
- It is remarkably potent (only a few dsRNA molecules per cell are required for effective interference).
- The interfering activity (and presumably the dsRNA) can cause interference in cells and tissues far removed from the site of introduction.

A possible mechanism of gene silencing is shown in Figure 4.9.

The dsRNA is recognized, and the DICER enzyme complex degrades it into short double-stranded fragments. The short regions then function as the recognition sites for any RNA that contains the same sequence to be degraded. This way all the RNA transcripts from any of the members of a gene family can be simultaneously silenced. Any resulting phenotype can then be attributed to the functioning of that gene family, but it will still need to be determined whether the family members contribute redundant functions or whether only one of the members of the gene family actually conditions the particular phenotype observed.

SUMMARY

Gene discovery involves a number of different phases, some of which can be carried out in parallel. Most of the discovery starts at the sequence level, either with genomic sequence or ESTs. The informatics analyses applied to the sequence data lead to tentative gene assignments. The confirmation of the function of these tentative expressed sequences (from genomic DNA) or actual expressed sequences is dependent on additional data. These data can come from comparative analyses with other organisms where the function of a similar sequence is known or from direct protein functional studies. Patterns of gene expression can also give indications of possible function. However, at present, it is not possible to go directly from nucleic acid sequence to protein function, and therefore gene identity, in any organism.

REFERENCES

Azpiroz-Leehan, R ,and K. A. Feldmann (1997) T-DNA insertion mutagenesis in *Arabidopsis*: going back and forth. *Trends Genet.* **13**, 152–156.

Chory, J., J. R. Ecker, S. Briggs, M. Caboche, G. M. Coruzzi, D. Cook, J. Dangl, S. Grant, M. L. Guerinot, S. Henikoff, R. Martienssen, K. Okada, N. V. Raikhel, C. R. Somerville. and D. Weigel (2000) National Science Foundation-Sponsored Workshop Report: "The 2010 Project" Functional Genomics and the Virtual Plant. A Blueprint for Understanding How Plants Are Built and How to Improve Them. *Plant Physiol.* **123**, 423–426.

Cogoni, C., and G. Macino (2000) Post-transcriptional gene silencing across kingdoms. *Genes Dev.* **10**, 638–643.

Colbert, T., B. J. Till, R. Tompa, S. Reynolds, M. N. Steine, A. T. Yeung, C. M. McCallum, L. Comai, and S. Henikoff (2001) High-throughput screening for induced point mutations. *Plant Physiol.* **126**, 480–484.

McCallum, C. M., L. Comai, E. A. Greene, and S. Henikoff (2000) Targeting Induced Local Lesions In Genomes (TILLING) for plant functional genomics. *Plant Physiol.* **123**, 439–442.

Oleykowski C. A., C. R. Bronson Mullins, A. K. Godwin, and A. T. Yeung (1998) Mutation detection using a novel plant endonuclease *Nucleic Acids Res.* **26**, 4597–4602.

Parinov, S., and V. Sundaresan (2000) Functional genomics in *Arabidopsis*: Large scale insertional mutagenesis complements the genome sequencing project. *Curr. Opin. Biotechnol.* **11**, 157–161.

Ramachandran, S., and V. Sundaresan (2001) Transposons as tools for functional genomics. *Plant Physiol. Biochem.* **39**, 243–252.

Tang G. L., B. J. Reinhart, D. P. Bartel, and P. D. Zamore (2003) A biochemical framework for RNA silencing in plants. *Genes Dev.* **17**, 49–63.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant, *Arabidopsis thaliana*. *Nature* **408**, 796–815.

Wang, E., R. Wang, J. DeParasis, J. H. Loughrin, S. Gan, and G. J. Wagner (2001) Suppression of a P450 hydroxylase gene in plant trichome glands enhances natural-product-based aphid resistance *Nature Biotechnol.* **19**, 371–374.

Walbot, V. (2000) Saturation mutagenesis using maize transposons. *Curr. Opin. Plant Biol.* **3**, 103–107.

Ware, D., P. Jaiswal, J. Ni, X. Pan, K. Chang, K. Clark, T. L. Clark, S. Schmidt, W. Zhao, S. Cartinhour, S. McCouch, and L. Stein (2002) Gramene: a resource for comparative grass. *Genomics Nucleic Acids Res.* **30**, 103–105.

CHAPTER

5

Control of Gene Expression

Overview

The control of gene expression can occur at many levels (Figure 5.1). The primary level involves the transcription of the DNA into RNA. There are many controls over the initiation of this event including the need for promoters and terminator sequences, for appropriate chromatin structure and the presence of enhancers or modulating sequences that affect the availability of a particular gene for transcription. Once transcribed the RNA then must be processed into a mature message and exported from the nucleus in a form that permits translation by the ribosomal machinery. The half-lives of various RNAs, and even their capability of being translated, differ widely. After translation the RNAs can either be recycled and translated again or can be targeted for degradation. The final points of control are at the protein level with protein modification events that affect the ultimate function of the protein product of the gene.

Not all genes are turned on at all times during the life cycle of a plant. Different genes are required for the completion of different steps in the developmental and sexual maturation of the plant and in response to the external environment (see Chapter 7). Two general types of control can be described:

- Temporal regulation—when a gene is only expressed at a specific time in development, for example, genes that are only expressed in the light or during flower development
- Spatial regulation—when a gene is only expressed in a specific location in the plant, for example, seed storage proteins or anther specific genes

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

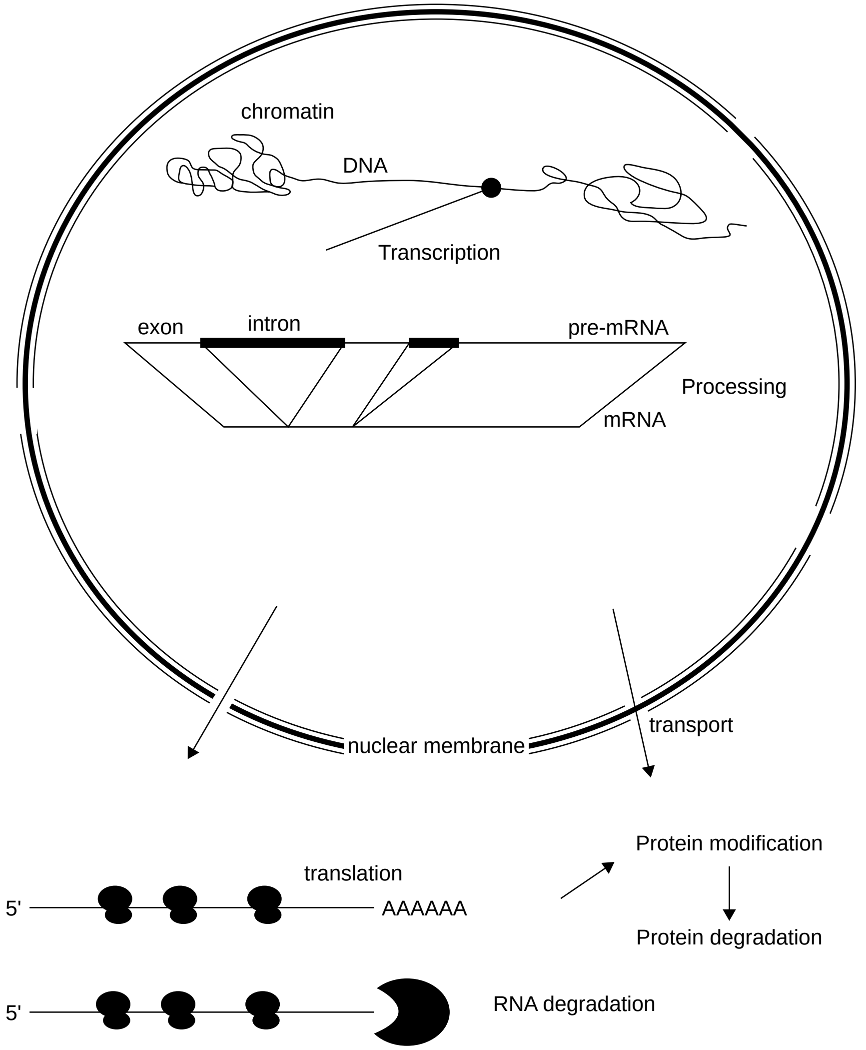


FIGURE 5.1. Points of control in gene expression. The region of the chromosome to be transcribed accumulates RNA polymerase. The nascent RNA is processed and then transported to the cytoplasm to be translated. The RNA is degraded or reused. The peptide is folded and/or modified to deliver the final product that is itself degraded at some point.

Gene expression is controlled by both the sequences surrounding the gene and the chromosome context in which the gene resides. The promoter elements can be complex containing sequences that control temporal and spatial regulation as well as enhancer and silencing activities. Added to these

elements, the role of 3′ sequences and introns in gene expression also must be considered. Chromatin variation brought about by the reversible modification of the core histones by acetylation, methylation, phosphorylation, ubiquitination, or ADP-ribosylation (Strahl and Allis, 2000) all have consequences for gene activation, gene repression, and chromosome replication. Epigenetic phenomena are alternative states of gene expression, molecular function, or organization specified by the same genetic instructions (the primary DNA sequence). Chromatin modifications including cytosine methylation and histone modifications are known to be involved in many such epigenetic phenomena (Habu et al., 2001).

Many of the experiments to determine the nature of control elements have been done by making transgenic plants. Unfortunately, high-efficiency homologous recombination is not currently available in plants so that it is not possible to place such control elements in their appropriate chromosomal context. This raises concerns because it is clear that the chromosomal context counts for the appropriate expression patterns to be reproduced, so that the expression patterns observed with transgenic plants may not accurately reflect the normal *in vivo* situation (Li et al., 2002).

Specific Promoter Sequences Are Required for Regulated Gene Expression

There are numerous types of promoters that can regulate gene expression. These promoters can be constitutive or inducible. A constitutive promoter is one that functions in all tissues under all conditions, whereas an inducible promoter is only activated in response to specific stimuli. A constitutive promoter may still drive differential levels of expression but should not be completely silent in any tissue. Table 5.1 lists a series of plant promoters and the expression patterns that are observed when they are used in transformation experiments (Lessard et al., 2002). The identification of such specific promoters is of use in controlling the expression of a transgene when it is necessary or desirable to target the expression of that transgene to specific organs, tissues, or developmental stages. In addition to using plant-specific promoters for this task, a number of chemically induced promoters have been developed to enable a controlled activation of constructs.

The Effect of Enhancer Elements on Gene Expression

An enhancer is a DNA sequence that greatly increases the expression of a gene in its vicinity. The sequence can reside upstream or downstream, its

TABLE 5.1. Examples of Different Types of Plant Promoters Used for Construction of Transgenes

Type of promoter	Name	Comments	Reference
Constitutive	Ubiquitin	Plant-origin promoter that drives high-level constitutive expression, but expression level during development may vary	Plesse et al., 2001
Constitutive	tCUP	Cryptic promoter from tobacco	Wu et al., 2003
Tissue-specific embryo promoter	β-Conglycinin	A well-characterized promoter that directs embryo-specific expression	Chen et al., 1986
Tissue-specific endosperm promoter	Opaque-2	These promoters show developmental regulation and expression	Rossi et al., 1997
Tissue-specific fruit	2A11	2A11 is a fruit-specific promoter derived from tomato	Van Haaren and Houck, 1993
Tissue-specific leaf	Lhcb3 promoter	Lhcb3 promoter is a light-regulated leaf-specific promoter from *Arabidopsis*	Ali and Taylor, 2001
Tissue-specific stolon	Stgan	Exclusive expression in stolons	Trindade et al., 2003
Pollen specific	lat52 promoter	Developmentally regulated but drives high expression during pollen maturation	Bate and Twell, 1998

Reprinted from *Metab. Eng.* **4**, Lessard et al. Manipulating gene expression for the metabolic engineering of plants, 67–79, Copyright 2002, with permission from Elsevier.

orientation is not fixed, and enhancers can still function when placed at some distance from the transcription start site. These elements do not control whether or not the promoter is active, but they do affect the level of expression. Their effects can be magnified if they are present in multiple copies. They can also have a greater enhancement when present in various combinations. For example, analysis of the tCUP promoter (Wu et al., 2003) demonstrated that when the long AT-rich region (between positions –350 and –161 bp relative to the transcription start site) was deleted a significant decrease in promoter activity was observed, whereas oligomerization of this region enhanced activity. A 21-bp sequence (TAGCCCCAA TTTCAAATTCAA) was also identified in a similar fashion and defined a novel cryptic constitutive enhancer element the activity of which was dependent on the AT-rich element itself.

As well as enhancing gene expression, some *cis*-acting regions can also silence the expression of adjacent genes. A bidirectional mannopine synthase (MAS) promoter contains elements that both enhance and silence gene expression (Guevara-Garcia et al., 1999; Figure 5.2). The bidirectional promoter controls the expression of two genes that encode enzymes for the synthesis of mannopine in plant tissues infected with *Agrobacterium tumefaciens*. The use of different MAS promoter regions in transformation experiments demonstrated that the regulatory properties of these divergent promoters result from the presence of orientation-dependent negative and positive regulatory regions. In particular, some of these elements have the unusual property of acting as enhancers in one orientation and as silencers in the other. This is an example of the complexity of the control and response elements over a short region adjacent to the gene.

Role of 5′ Sequences in Gene Expression

Certain 5′ leader sequences have been shown to enhance the expression of transgenes. Both viral and endogenous plant 5′ untranslated leader sequences have effects. The ability of 5′ leader sequences to enhance gene expression has been used to design more efficient transgene expression systems for plants (Groenewald et al., 2000). However, the 5′ untranslated leader sequences from dicots appear to work better in dicots than those from monocots and vice versa.

Role of 3′ Sequences in Gene Expression

There is evidence that the transcription-enhancing effects of the 3′ coding sequence for transgenes may be related to the effectiveness of transcription termination (Ali and Taylor, 2001; Bieri et al., 2002). As with many of the

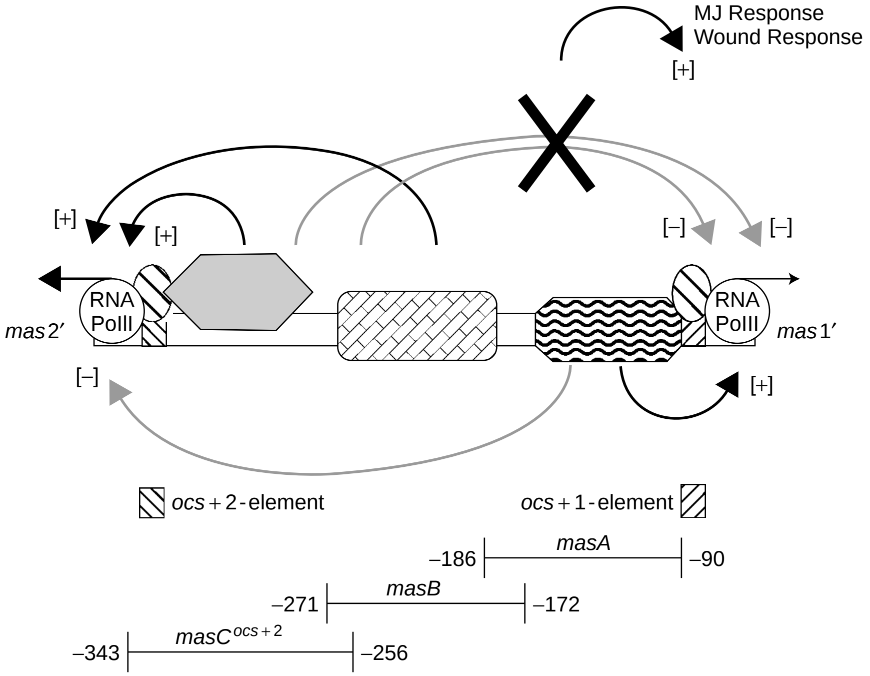


FIGURE 5.2. Functional model of MAS bidirectional promoter activity. The capacity of MAS subdomains to bind specific nuclear proteins was defined by footprinting analysis. The relative positions of ocs-like elements and MAS promoter fragments with enhancer or silencer activity are indicated. The ocs-binding factors at the ocs-like elements and the putative locations of the plant nuclear proteins that bind to the MAS promoter are depicted according to Ni et al. (1996). The relative strength and transcriptional orientation of the mas2 and mas1 promoters are indicated by the horizontal arrows. The orientation-dependent function of the each fragment is indicated by the curved arrows. Positive regulatory effects are indicated by [+] signs and heavy curved arrows, whereas negative regulatory effects are indicated by [–] signs and stippled curved arrows. The suppressor effect that wounding and methyl jasmonate (MJ) have on the silencer properties of masB and masCocs+2 regions is represented by the large X. (Reprinted with permission from Guevara-Garcia et al., 1999. Copyright Springer-Verlag GmbH & Co. 1999).

transcription signals, the generality of these conclusions may not be equally applicable across the whole plant kingdom.

Role of Introns in Gene Expression

Introns have been shown to be important for improving transgene expression (Ibrahim et al., 2001) and can also play an important role in

determining the tissue specificity of gene expression. The enhancement can be more than 100-fold (Maas et al., 1991; Zhang and Singh, 1994), but is usually in the range of 2- to 10-fold. The effect is generally larger in monocots than dicots. The intron must be within the transcribed portion of the gene for the intron-mediated enhancement (IME) to occur, because there is no effect if the intron is placed upstream of the promoter. Therefore, the expression effect is not due to any enhancer-like activity. A number of mechanisms have been proposed to explain this so-called intron-dependent gene expression:

- Spliceosome formation plays a role in stabilizing transcripts in the nucleus (Luehrsen et al., 1994)
- The presence of negative or positive *cis*-acting sequences in genes determines whether or not introns are necessary for proper expression. Any *cis*-acting elements in a particular gene, in the absence of an intron, act to impair expression of genes (Rethmeir et al., 1998). Therefore, introns do not necessarily enhance gene expression but act to overcome repression of the normal expression levels of the genes in question.

Results from use of either of the first two introns of the *Arabidopsis* tryptophan pathway gene PAT1 showed a roughly fivefold elevation of mRNA accumulation from a PAT1:β-glucuronidase (GUS) fusion without affecting the rate of PAT1:GUS transcription. An analysis of the modification of such constructs eliminated the possibility that the intron stimulated mRNA accumulation via a unique RNA-stabilizing sequence or through the completed act of splicing. However, the results were consistent with a possible role for redundant intron sequence elements or an association of the pre-mRNA with the spliceosome.

As has been described with other control sequences, any particular intron does not work equally well when introduced into monocots and dicots. Therefore, the requirements for correct splicing may not be the same in these two groups. Introns that stimulate expression in monocots include those from the maize *Adh1*, *Sh1*, *Bz1*, *Hsp82*, actin, and *GapA1* genes and the rice *salT*, *Act1*, and *tpi* genes (cited in Rose and Beliakoff, 2000). Dicot introns that elevate expression include those from the petunia *rbcS* gene *SSU301*, the potato ST-LS1 gene, and the *Arabidopsis UBQ3*, *UBQ10*, *PAT1*, *atpk1*, A1 EF-1a, and *At eEF-1*b genes (cited in Rose and Beliakoff, 2000).

Conserved Sequences in Eukaryotic Promoters

The ability to predict gene structure from genomic sequences would be easier with knowledge of the important sequences surrounding the start of translation. Therefore, the ability to identify plant promoters will aid gene

annotation efforts as well as improving the understanding of the regulation of gene expression. A knowledge of the features of promoter elements surrounding putative genes would aid in the prediction of the type of conditions and/or tissues in which the putative gene might be expressed, thereby directing the search for confirmation of the gene and its structure to the appropriate target tissue. The selection of the transcription start site, specificity, and rate is controlled by elements of the promoter that can be thousands of nucleotides from the actual transcription start site. A number of databases that contain *cis*-acting sequences are available. A new plant specific database, PlantPromDB, release 2001.01, contains 305 entries including 71, 220, and 14 promoters for RNA polymerase II from monocot, dicot, and other plants, respectively (Shahmuradov et al., 2003). A second database, Plant CARE (Lescotet et al., 2002) contains plant *cis*-regulatory elements, enhancers, and repressors. This database currently contains 417 *cis*-regulatory elements, 150 of which are from monocots, 263 from dicots, and 4 from conifers derived from 160 individual promoters from higher-plant genes.

Trans-Acting Factors Control Gene Expression

Any consideration of the sequences that comprise promoters must also include the proteins that recognize these sequences and their binding to the DNA sequences that surround the transcription start site. Thus the promoters are the *cis*-acting regions that are required for gene expression whereas the *trans*-acting factors are usually proteins that bind to these *cis*-acting sequences to effect the controlled transcription.

Transcription factor genes make up a substantial fraction of all eukaryote genomes. These genes can be grouped into gene families according to the type of DNA binding domain that they encode. These families may contain large numbers of members and include functional redundancy within a family. The transcription factors can interact both among themselves and with other proteins in their role in which they control expression of the genome at the transcriptional level. The major families of transcription factors in *Arabidopsis* are shown in Table 5.2 (from Reichmann and Ratcliffe, 2000).

The use of microarrays for transcriptional profiling will also be important in the understanding of both the interaction of transcription factors with the DNA and the conservation of *cis*-acting elements among coordinately regulated genes. As the upstream and downstream sequences of coordinately regulated genes are identified and placed in queryable databases, the conservation of either sequence or motif will become easier to observe and better understood. This information will then be useful in the design of new transformation vectors.

TABLE 5.2. Major Families of *Arabidopsis* Transcription Factors

Gene family	Estimated number of genes in the *Arabidopsis* genome	Gene family functions	Genetically characterized *Arabidopsis* factors
MYB	180	Secondary metabolism, cellular morphogenesis, signal transduction in plant growth, abiotic and biotic stress responses, circadian rhythm, dorsoventrality	AtMTB2, ATR1, CCA1, CPC, GL1, LHY, WER
AP1/EREBP	150	Flower development, cell proliferation, secondary metabolism, abiotic and biotic stress responses, ABA response, ethylene response	ABI4, ANT, AP2, CBF1-3/DREB1 A-C, DREB2A, ERF1
NAC	105	Development, pattern formation, organ separation	CUC2, NAP
bHLH/MYC	100	Anthocyanin biosynthesis, light responses, flower development, abiotic stress responses	PIF3
bZIP	100	Seed storage gene expression, photomorphogenesis, leaf development, flower development, defense response, ABA response, gibberellin biosynthesis	ABI5, HY5, PAN

TABLE 5.2. Continued

Gene family	Estimated number of genes in the *Arabidopsis* genome	Gene family functions	Genetically characterized *Arabidopsis* factors
HB	90	Development (leaf, root, internode, and ovule), stem cell identity, cell differentiation, growth response, anthocyanin accumulation, cell death	ANL2, ANTHB2, BEL1, GL2, KNAT1, REV, STM, WUS
ZC_2H_2	85	Flower development, flowering time, seed development, root nodule development	FIS2, SUP
MADS	80	Flower development, fruit development, flowering time, root development	AG, AGL15, ANR1, AP1, AP3, CAL, FLC, FUL, Pl, SEP1, SEP2, SEP3, SHP1, SHP2, SOG1, SVP
WRKY	75	Defense response	
ARF-Aux/IAA	42	Auxin responses, development, floral meristem patterning	AXR2, AXR3, ETT, MP, NPH4, SHY2
Dof	41	Seed germination, endosperm-specific expression, carbon metabolism	DAG1

Reprinted from *Curr. Opin. Plant Biol.* **3**, Riechmann and Ratcliffe, A genomic perspective on plant transcription factors, 423–434, Copyright 2000, with permission from Elsevier.

mRNA STABILITY

In addition to regulation at the transcriptional level, posttranscriptional mechanisms also play a fundamental role. The control of mRNA stability is one of the forms of posttranscriptional regulation, and the stability of a particular mRNA determines, to a great extent, its steady-state levels. The half-lives of mRNAs can vary over a wide range. General structural elements found at the ends of the mRNA, as well as specific sequence elements located within a transcript, can all contribute to the overall stability of the message. Specific sequence elements involved in the control of mRNA stability include:

- Downstream element (DST) identified from the unstable small auxin-up RNA (*SAUR*) transcripts
- The light-responsive element
- The 3′ UTR involved in sugar repression

Although it is clear that unstable mRNAs contain instability sequences, no discrete stabilizing determinant has been demonstrated to be responsible for the long half-life of an extremely stable transcript in plant systems (Gutierrez et al., 1999).

CHEMICALLY REGULATED GENE EXPRESSION IN PLANTS

Understanding the developmental and physiological responses in plants can be facilitated by the targeted expression or suppression of specific genes. This is frequently achieved through the use of transgenic plants, and the precise control of the regulation of the expression of the introduced gene in both time and space is necessary. In this context an expression cassette should have two features:

1. The inducer must be specific and nontoxic.
2. The target promoter should have a very low noninduced activity and a high induced activity.

The characteristics of several chemically inducible systems have been reviewed (Padidam, 2003) and are included in Table 5.3 from this source.

EFFECTS OF CHROMATIN STRUCTURE

It is clear that chromatin (the proteinaceous material that together with DNA comprises the chromosomes) structure is not just a mechanism for conveniently packaging the DNA within the nucleus. The chromatin is a dynamic

TABLE 5.3. Strengths and Limitations of Chemically Regulated Systems Developed for Plant Gene Expression

System	Tested in	Strengths	Limitations
Tetracycline-inducible TetR	Tobacco, tomato, potato, BY2 cells, SR1 cells	Low amounts of inducer are sufficient for activation, inducer readily taken up by cells	Leaky expression, high concentrations of TetR are required, short half-life of inducer, does not work in *Arabidopsis*
Tetracycline-inactivatable tTA	*Arabidopsis*, tobacco, NT1 cells	Target promoter can be shut off efficiently, turnover of transgene can be assessed, lower basal levels compared with TetR	Plants must be maintained with tetracycline to turn off gene expression, negatively controlled system
Dexamethasone-inducible GR fusions	*Arabidopsis*, tobacco	GR::transcription factor fusion proteins identify immediate target genes, posttranscriptional activation	Only suitable for transcription factors, inducer toxic in some cases, induction of defense-related genes in *Arabidopsis*
Dexamethasone-inducible GVG	*Arabidopsis*, tobacco	Dexamethasone easily permeates plants cells and can be applied by various routes	Inducer toxic in some cases, induction of defense-related genes in *Arabidopsis*, slow off rate
Estradiol-inducible ER–C1	Black Mexican sweet corn cells	Relatively low levels of estradiol are required, no apparent toxic effects	Not tested in transgenic plants, not suitable for plants with phytosteroids
Estradiol-inducible XVE	*Arabidopsis*, tobacco, BY2 cells	No apparent toxic effects, low basal level and high inducible levels	Not suitable for plants with phytosteroids (soybean), not suitable for field use

Dex-inducible and tet-repressible TGV	*Arabidopsis*, tobacco, BY2 cells	Dual control, quick shut off	Inducer toxic in some cases, defense-related genes may be induced
Tebufenozide-inducible GVHvEcR	Tobacco	Safe inducer, suitable for field application	High basal activity, foliar uptake of inducer is poor
Methoxyfenozide-inducible GVCfEcR	*Arabidopsis*, tobacco, BY2 cells	Low basal and high inducible levels, safe inducer, suitable for field application	Foliar uptake of inducer is poor, turn off is slow
Methoxyfenozide-inducible GVOnEcR	Maize	Safe inducer, inducer moves systemically, suitable for field application	Foliar uptake of inducer is poor
Ethanol-inducible AlcR	*Arabidopsis*, tobacco, potato	Inducer inexpensive and biodegradable, rapid reversible induction, suitable for field application	Volatile inducer, induction can triggered inadvertently, induction due to anoxia, inducer cannot be used for more than 2 d
Copper-inducible ACEI	*Arabidopsis*, tobacco, root nodules, BY2 cells	Ease of application of inducer, simple to use, inducer inexpensive	Induced levels are low, long exposure of inducer is toxic, does not work in BY2 or protoplasts derived from tobacco leaves, expression is variable in *Arabidopsis*
Benzothiadiazole-inducible PR-1a	Tobacco	Inducer not phytotoxic, long-lasting response, suitable for expression of disease resistance genes in field	Low induction levels, promoter also responds to oxidative stress and endogenous salicylic acid signals, inducer activates native genes
Safener-inducible In2-2	*Arabidopsis*, BY2 cells	Suitable for field application, inducer is an agrochemical	Inducer causes growth abnormalities, constitutive expression in roots, promoter responsive to other chemicals

Reprinted from *Curr. Opin. Plant Biol.* **6**, Padidam, Chemically regulated gene expression in plants, 169–177, Copyright 2003, with permission from Elsevier.

material that changes as the internal cellular environment alters, whether in response to the external environment or to endogenous changes. A key requirement for the expression of genes in chromosomes is that chromatin be remodeled (i.e., "opened") in such a way that transcriptional activator proteins and RNA polymerases can have access to the DNA, permitting the assembly of a transcription complex that then transcribes the gene into messenger RNA.

The nucleosome, in which a loop of DNA about 150 nucleotides long is wrapped around a core of histone proteins, is the basic unit of chromatin. This histone core consists of an octamer containing two copies of each of the four histones H2A, H2B, H3, and H4. The nucleosomes are packaged in successively higher orders of coiling in the chromosomes. The changes in chromatin structure are mediated through modifications to the histones that include acetylation, phosphorylation, ubiquitination, and ADP-ribosylation of particular amino acid residues. The role of the *Arabidopsis* gene DDM1 (Decrease in DNA Methylation 1) indicates that the chromatin modification can also be intimately involved with DNA methylation (Singer et al., 2001).

The change in chromatin state, in itself, does not necessarily result in the transcription of the region but simply alters the accessibility of the region. Therefore, it is still necessary for the appropriate *trans*-acting factors, such as transcription factors, to be present for the gene transcription to occur. Conversely, however, the lack of chromatin remodeling can prevent gene transcription even in the presence of the appropriate *trans*-acting factors.

Matrix attachment regions (MARs) that flank some plant genes may also be necessary for the correct regulation of those genes. For example, the MARs in addition to the downstream region and introns from the heat shock cognate 80 gene (HSC80) of tomato are necessary for the efficient expression of HSC80 transgenes (Holmes-Davis and Comai, 2002). The MARs are identified by their ability to bind the nuclear matrix and are AT rich. They also include the topoisomerase II consensus sequence. Two possible methods by which MARs may exert their influence are:

- By anchoring the ends of chromatin loops to the nuclear matrix and resulting in an independent chromatin domain
- MARs might be important for interactions between the activating complexes and the DNA

In the first case MARs should be nonspecific and function to activate most genes, whereas in the second case the effects are more likely to be gene specific. These two effects may not be mutually exclusive, nor is it necessary that all MARs function in the same manner. Support for the role of MARs in the interaction with gene-specific elements for the induction of transcriptional competence and therefore a limited role in heterologous contexts has been reported (Holmes-Davis and Comai, 2002).

TRANSLATIONAL CONTROL

Translation is a potential step in the regulation of gene expression. However, this stage is less well understood than the control of transcription. For example, some plant viral RNAs are more efficiently translated than the endogenous host mRNAs and this enhancement is due to the presence of translational enhancers in the 5′ leader sequence.

SUMMARY

An overall view of the regulation of transcription should ultimately emerge from the whole genome analysis that can be performed in *Arabidopsis* and rice with the complete genome sequence. From these analyses the *cis*-acting elements that are necessary for the ultimate control of gene expression will become apparent, as well as the characterization and understanding of all interactions between transcription factors, other proteins, and the *cis*-acting elements. As with all of the genome-wide studies, large amounts of data will need to be correlated and integrated in order for the overall picture to become clearer.

REFERENCES

Ali, S., and W. C. Taylor (2001) The 3′ non-coding region of a C4 photosynthesis gene increases transgene expression when combined with heterologous promoters. *Plant Mol. Biol.* **46**, 325–333.

Bieri, S., I. Potrykus, and J. Fütterer (2002) Geminivirus sequences as bidirectional transcription termination/polyadenylation signals for economic construction of stably expressed transgenes *Mol. Breeding* **10,** 107–117.

Groenewald, J.-H, N. F. Hiten. and F. C. Botha (2000) The introduction of an inverted repeat to the 5′ untranslated leader sequence of a transgene strongly inhibits gene expression. *Plant Cell Rep.* **19**, 1098–1101.

Guevara-Garcia, A., J. Lopez-Bucio. and L. Herrera-Estrella (1999) The mannopine synthase promoter contains vectorial cis-regulatory elements that act as enhancers and silencers. *Mo*l. *Gen. Genet.* **262**, 608–617.

Gutiérrez, R. A., G. C. MacIntosh. and P. J. Green (1999) Current perspectives on mRNA stability in plants: multiple levels and mechanisms of control. *Trends Plant Sci.* **4**, 429–438.

Habu, Y., T. Kakutani, and J. Paszkowski (2001) Epigenetic developmental mechanisms in plants: molecules and targets of plant epigenetic regulation. *Curr. Opin. Genet. Dev.* **11**, 215–220.

Holmes-Davis, R., and L. Comai (2002) The matrix attachment regions (MARs) associated with the Heat Shock Cognate 80 gene (HSC80) of tomato represent specific regulatory elements *Mol. Genet. Genomics* **266**, 891–898.

Ibrahim, A. F. M., J. A. Watters, G. P. Clark, C. J. R. Thomas, J. W. S. Brown, and C. G. Simpson (2001) Expression of intron-containing Gus constructs is reduced due to activation of a cryptic 5′ splice site. *Mol.Genet. Genomics* **265**, 455–460.

Lessard, P. A., H. Kulaveerasingam, G. M. York, A. Strong, and A. J. Sinskey (2002) Manipulating gene expression for the metabolic engineering of plants. *Metab. Eng*. **4**, 67–79.

Lescot, M., P. Dehais, G. Thijs, K. Marchal, Y. Moreau, Y. Van De Peer, P. Rouze, and S. Rombauts (2002) Plantcare, a database of plant *cis*-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res*. **30**, 325–327.

Luehrsen, K.R., S. Taha, and V. Walbot (1994) Nuclear pre-mRNA processing in higher plants. *Prog. Nucl. Acid Res. Mol. Biol*. **47**, 149–193.

Li, G., T. C. Hall, and R. Holmes-Davis (2002) Plant chromatin: Development and gene control. *BioEssays* **24**, 234–243.

Maas, C., J. Laufs, S. Grant, C. Korfhage, and W. Werr (1991) The combination of a novel stimulatory element in the first exon of the maize *Shrunken-1* gene and the following intron 1 enhances reporter gene expression up to 100-fold. *Plant Mol. Biol.* **16**, 199–207.

Ni, M., D. Cui, and S. B. Gelvin (1996) Sequence-specific interactions of wound-inducible nuclear factors with mannopine synthase 2′ promoter wound-responsive elements. *Plant Mol. Biol.* **30**, 77–96.

Padidam, M. (2003) Chemically regulated gene expression in plants. *Curr. Opin. Plant Biol.* **6**, 169–177.

Rethmeier, N., J. Seurinck, M. Van Montagu, and M. Cornelissen (1997) Intron-mediated enhancement of transgene expression in maize is a nuclear, gene-dependent process. *Plant J*. **12**, 895–899.

Riechmann, J. L., J. Heard, G. Martin, L. Reuber, C.-Z. Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J. Z. Zhang, D. Ghandehari, B. K. Sherman, and G.-L. Yu (2000) *Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* **290,** 2105–2110.

Riechmann, J. L., and O. J. Ratcliffe (2000) A genomic perspective on plant transcription factors *Curr. Opin. Plant Biol.* **3**, 423–434.

Rose, A. B., and J. A. Beliakoff (2000) Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing. *Plant Physiol.* **122**, 535–542.

Shahmuradov, I. A., A. J. Gammerman, J. M. Hancock, P. M. Bramley, and V. V. Solovyev (2003) Plantprom: a database of plant promoter sequences. *Nucleic Acids Res.* **31**, 114–117.

Singer, T., C. Yordan, and R. A. Martienssen (2001) Robertson's Mutator transposons in *A. thaliana* are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). *Genes Dev.* **15**, 591–602.

Strahl, B. D, and C. D. Alliss (2000) The language of covalent histone modifications. *Nature* **403**, 41–45.

Wu, K. Q., M. Hu, T. Martin, C. M. Wang, X. Q. Li, L. N. Tian, D. Brown, and B. Miki (2003) The cryptic enhancer elements of the Tcup promoter. *Plant Mol. Biol.* **51**, 351–362.

Zhang, B., and K. B. Singh (1994) ocs element promoter sequences are activated by auxin and salicylic acid in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **91**, 2507–2511.

CHAPTER

6

FUNCTIONAL GENOMICS

OVERVIEW

Cellular processes are controlled at various levels. The basic information is encoded by the genome, which essentially is identical in every cell of the organism, irrespective of the stage of development or environment. Control of the resultant bioactivity stemming from this invariant information can occur at a number of points, as shown in Figure 6.1.

The information contained in the DNA, the genome, can be transcribed into RNA, but only a small part of the information in the DNA is ever transcribed. The sum of all of the regions of genome that are transcribed is collectively known as the transcriptome and comprises all of the sequences that make up the complete RNA component of the organism. Because not all of the DNA is transcribed, the transcriptome is less complex than the genome. The products of the translation of the transcriptome generate the proteome. However, because some of the RNA molecules are never translated into proteins, the complexity of the proteome is less than that of the transcriptome. However, this complexity can be increased through posttranslational modifications of the proteins, giving the potential for more than one important product to be derived from a given message. These proteins then function in the synthesis of primary and secondary metabolites, the total of these making up the metabolome. The combination of the RNAs, proteins, and metabolites, functionally integrated together, results in the biological activity of a cell or tissue, and information on all of these processes is necessary for a complete understanding of plant form, function, and development.

DNA sequence information has been, and is continuing to be, gathered both from genomic DNA and cDNA studies. However, all of this sequence information is just a first step in understanding the coordinated control of

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

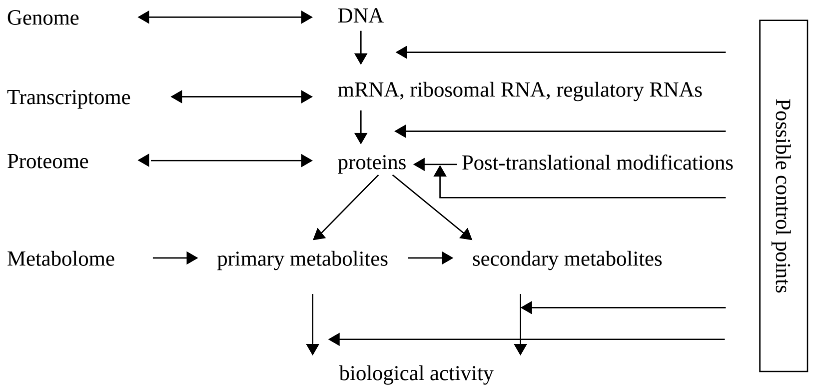


FIGURE 6.1. Levels of control for cellular processes (Adapted from Jacobs, D. I., R. van der Heijden, and R. Verpoorte (2000) Proteomics in plant biotechnology and secondary metabolism research. *Phytochem. Anal.* **11**, 277–287).

gene expression, as well as understanding which of these potential genes are actually expressed in a functional fashion. As mentioned in Chapter 2, the annotation of the genomic sequence refers to the processes (both manual and computer based) by which various regions of the genome are classified (genic regions, nontranscribed regions, transposable elements, conserved control regions). However, this analysis cannot address the spatial and temporal patterns of expression of those genes, or even whether those sequences are ever expressed. The data from ESTs and full-length cDNA clones directly identify the genomic sequences transcribed into RNA. However, even these data do not unambiguously identify those RNAs that ultimately influence the phenotype of the organism. Three possible fates for an RNA molecule are:

- It is translated into a protein.
- It functions as an RNA without any subsequent modification.
- It is degraded without ever actually performing a specific known function.

Therefore, to understand the functioning of the cell, additional information concerning transcription and translation (if appropriate) of any of the putative genes and the subsequent activity of any protein products must be obtained.

The three levels of expression, RNA, proteins, and products are covered in this chapter. Useful definitions are:

1. Expression profiling covers the methods used to acquire the information related to the RNA components of the cell.

2. The whole complement of sequences that is ever transcribed is also termed the transcriptome.
3. The identification and modification of proteins is now usually identified as proteomics, whereas the characterization of the final products of protein activity is sometimes referred to as metabolomics.
4. The term "proteome" refers to the complete cell protein complement, including unmodified and posttranslationally modified proteins.
5. "Metabolomics" refers to the methods used to characterize the metabolism of the cell.
6. "Metabolome" refers to the suite of possible metabolites within a plant.

Expression Profiling

The development of high-throughput methods has certainly changed the way in which the coordination of the expression of genes can be studied. The traditional method for determining where and when a gene is expressed was the Northern blot, which involved the hybridization of a single labeled probe to an RNA target and measurement of both the size of the band and the intensity of the signal. With the advent of expression profiling using microarrays, the level of expression of many thousands of genes in various tissues of plants grown under numerous conditions can be rapidly undertaken. As is the case with the explosion of genomic sequence data, the ability to manage all of the data that are generated is one of the challenges arising from these high-throughput methods. The design of these expression-profiling experiments is important so that the data that are generated can be analyzed in a meaningful fashion. In general, significant differences in the expression of genes that are highly expressed, and/or have large changes in their expression, will be apparent under most experimental designs. Those genes that are expressed at very low levels, or have small changes in expression that may be very meaningful, pose additional problems. The experimental design is of great importance for such variation to be statistically significant.

Expression profiling is essentially the identification of all of the RNAs that are present in a specific tissue sample at a particular time. Therefore, characterization of the RNA populations in various tissues can be a window on the changes in the underlying biochemical processes that are occurring. The development of a whole range of techniques that allow many, or all, of the RNAs in a sample to be visualized simultaneously means that a global expression profile showing the relative abundances of the vast majority of RNAs can now be undertaken. The various techniques by which this profiling can be performed include:

- Microarray analysis
- EST sequencing
- Serial analysis of gene expression (SAGE)
- Massively parallel signature sequencing (MPSS™)
- Differential display

These techniques can be divided into two types. The first type is where the estimate of expression is based on a hybridization signal intensity such as that derived from a Northern blot or a microarray The relative intensity of the signals, rather than the absolute value of the signal, is used. The second type is based on a direct count of the number of each of the RNAs that are present in the sample, as is done when using ESTs, SAGE, and MPSS™.

DNA Microarrays for Expression Analysis

DNA microarrays, frequently referred to as "chips", have resulted in a revolution in the analysis of gene expression (Lockhart and Winzeler, 2000). The expression levels for many thousands of genes can be simultaneously determined with these microarrays. A flowchart for the design of such an experiment is shown in Figure 6.2.

The first stage is the design of the array itself. Two different types of arrays are used. The first type of array is one where fragments from either genomic clones or cDNAs are amplified by PCR and then spotted on to an appropriate substrate. The second type is where short oligonucleotides are designed, usually from genomic sequence information, synthesized, and then attached to the substrate (Lipshutz et al., 1999; Kane et al., 2000). Each of these arrays gives slightly different data sets, although both forms are used extensively. The actual application is frequently the determining factor in the choice of which type to use. Therefore, the first consideration is the actual design of the chip and what sequences, or oligonucleotides, should be included. When little is known about what may be happening in any of the comparisons, the chip with the most diverse collection of potential genes on it is likely to be the most useful. As data begin to accumulate, a more selective array can be designed to answer more specific questions. Once the array is manufactured, hybridizations with the labeled probe can be performed. The data must be analyzed for patterns of expression that may change with the various treatments, time points, developmental stages, or other variables.

The scheme outlined in Figure 6.2 is for a hypothetical comparison of the genes expressed in healthy leaves compared with diseased leaves. The sampling of the diseased tissue would need to be done at various times after the initial infection to determine the time course of the infection process. An alternative experiment could be the comparison of the expression in leaves

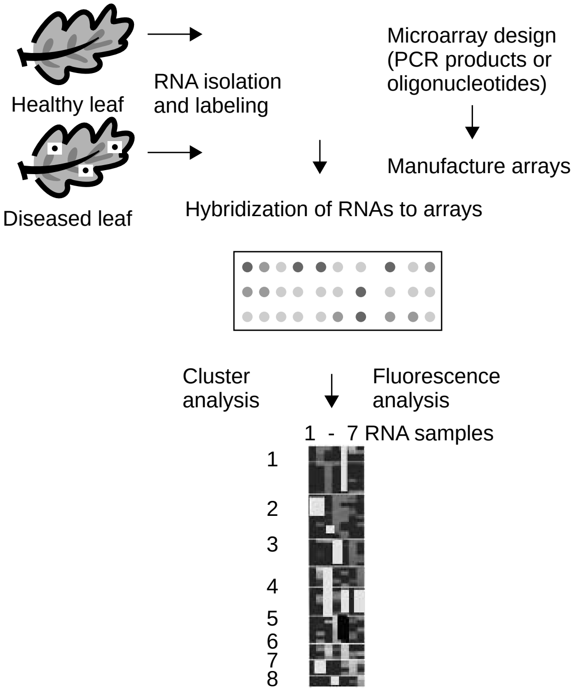


FIGURE 6.2. Outline of an expression profiling experiment using microarrays. RNA is extracted from healthy and diseased leaves and labeled. The labeled RNAs are hybridized to the microarray, and the fluorescence is detected. The data are processed and analyzed. DNAs on the microarray that have similar expression patterns are clustered and displayed. In this example 8 groups that have different patterns of expression for the RNA samples (1–7) are shown.

from a susceptible and a resistant variety at the same times after the initial challenge. The RNAs are extracted for each of the samples, and each of the RNAs is divided into two. One half will be labeled with one fluorochrome (e.g., Cy3) and the other half labeled with a different fluorochrome (Cy5). Then the microarray is hybridized with a mixture of the two samples, sample 1 labeled with Cy3 and sample 2 labeled with Cy5. The reverse hybridization to another of the microarrays, sample 1 labeled with Cy5 and sample 2 labeled with Cy3, will account for variation in labeling efficiencies and RNA quality. After hybridization and washing, the microarrays are scanned at two wavelengths and the signals are combined. If the signals from two fluorochromes are false-colored red and green, then when the hybridization is stronger with one of the samples, the spot will appear red or green. If the intensity of binding of both labeled RNAs is the same, then the spot on the microarray will appear to be yellow. Spots that have similar patterns of

expression across a range of samples are grouped together, allowing a visual representation and identification of the genes whose expression appears to be coordinately controlled.

Microarray expression analysis is limited by a number of factors:

- The sensitivity of the quantity of RNA that is hybridized to the chip
- The background intensity may overwhelm weak signals from lowly expressed transcripts because of nonspecific binding to the chip. This value can be estimated from standards that are included in the original microarray design and the signal corrected.
- The ease of detection of the differential expression of various members of a gene family or the detection of alternative splicing is dependent on the microarray design. Microarrays that consist of oligonucleotides are more effective in highlighting such differences (Grabowski, 2002; Modrek and Lee, 2002) (Figure 6.3).

Microarray experiments also require estimates of error and variability between samples. Therefore, replicates are needed to account for both biological and experimental variability (Churchill, 2002).

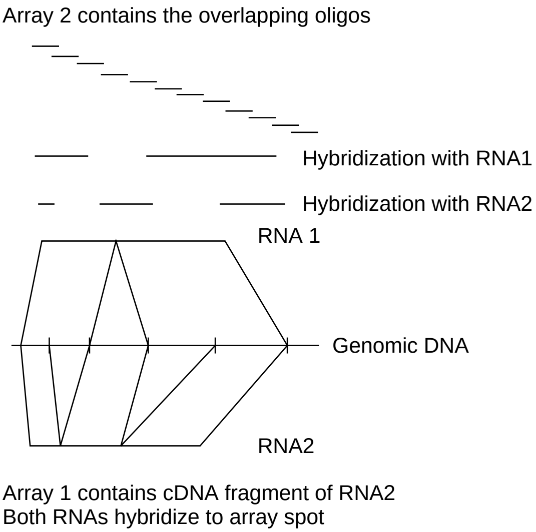


FIGURE 6.3. Detection of alternative splicing with oligo-based microarrays compared with PCR amplification products spotted onto the arrays. The genomic DNA transcript can be alternatively spliced into either of the RNAs 1 or 2. If RNA 2 has been cloned as a cDNA and the insert from that cDNA amplified and placed on the array, both RNAs 1 and 2 will hybridize to the spot. However, if a series of overlapping oligos are placed on the microarray, then the two patterns of hybridization for the two processed RNAs will be very different and distinguishable.

- Variability arising from the array manufacture and nonspecific binding and labeling can be estimated by the use of multiple positions within the array of the same sample, the inclusion of known standards, and the use of dye interchange for labeling the samples.
- The biological variation between samples must be controlled to make valid comparisons of the expression patterns. Multiple independent extractions from the tissues are necessary to estimate for diverse tissues or treatments.

These microarray hybridization studies are exceptionally useful for comparative experiments where the level of expression of a large number of genes must be compared under different conditions. Because the design of the array itself requires some prior selection of the sequences to be included, this technology in itself is not a method for finding new genes. Although the methodology is good for the comparison of expression levels of different genes, it does not lend itself to the absolute determination of the number of copies of a particular RNA that is present in the probe. The hybridization signal cannot be used to determine the absolute levels of each RNA in a sample.

COUNTING RNA MOLECULES

The direct detection of RNA sequences by ESTs, SAGE (Powell, 1998), and MPSS™ (Brenner, 2000) means that these methods give a quantitative value for the differential expression of each message without the need for standardization or repetition of every experiment. This results from the fact that these methods make a direct assessment of the relative abundance of each transcript from the number of times that that transcript appears in the collection provided that the extraction and construction of the sampled populations have not themselves introduced any biases.

EST ANALYSIS OF GENE EXPRESSION

As described in chapter 4, one way of finding genes is simply to sequence cDNA clones and analyze these transcripts. These ESTs are a sample of the sequences that are present in that particular cDNA library. Analysis of the ESTs from libraries made with various tissue sources will highlight any differential gene expression found in these tissues simply from the relative abundance of each of the sequences generated. The more deeply any particular cDNA library is sequenced, the more accurate the count of the number of copies of each transcript per cell will become. With continued sequencing from the library even rarely expressed transcripts can eventually be discovered. However, a representation of every sequence that is present in the cDNA library would be very costly to generate and would also result in a

highly redundant sequencing effort. In comparison to other digital methods such as SAGE and MPSS™, ESTs reveal additional information rather than only counting the relative abundance of RNAs. They reveal new genes, splice sites that will aid in computer-based gene annotation, and termination sites. However, because they are not full-length cDNAs and are the result of single-pass sequencing efforts, the discrimination between the different members of multigene families will not be very efficient and will generate much redundant sequence.

SAGE Analysis

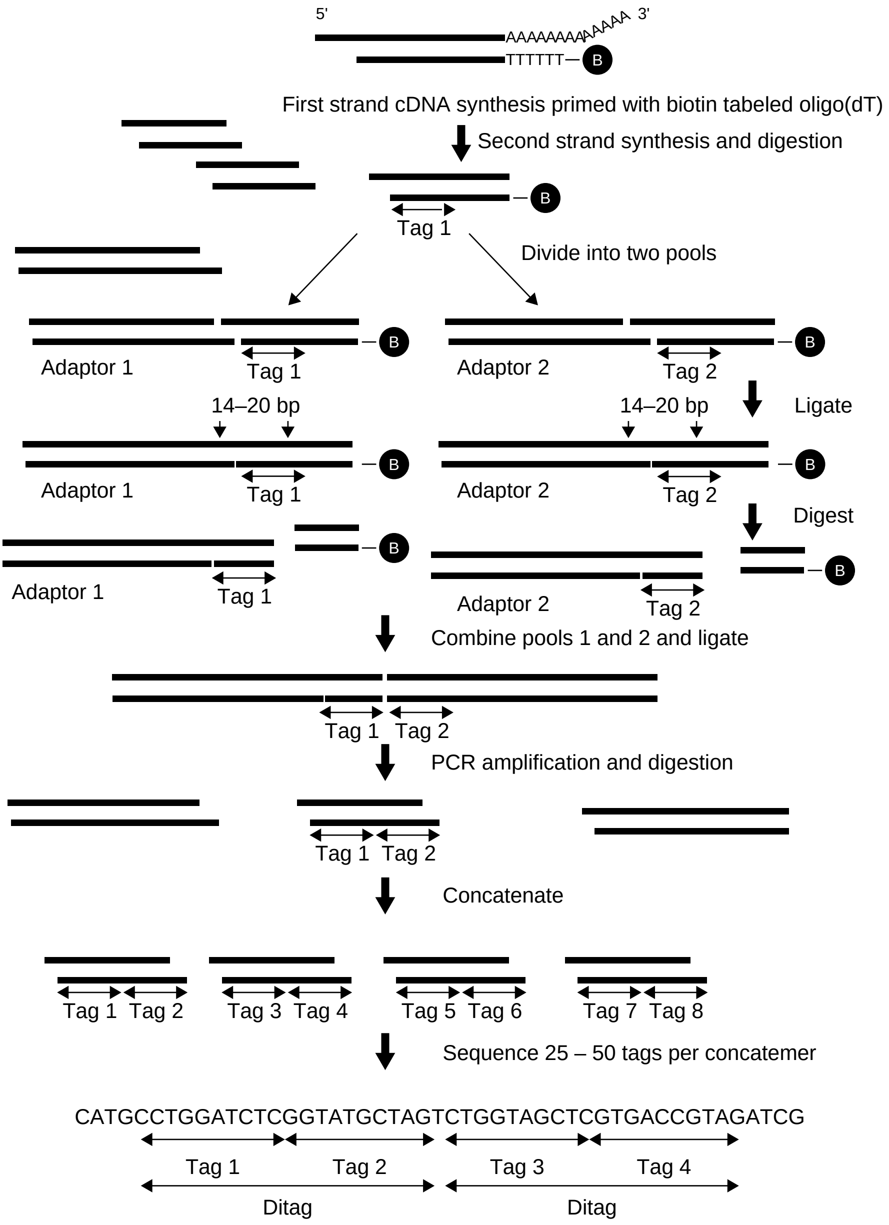
Serial analysis of gene expression (SAGE) is a sequence-based approach that identifies which genes are expressed and quantifies the level of their expression (Velculescu et al., 1995; Madden et al., 2000). It is essentially a modification of the process of generating ESTs. The usefulness of the method and its advantages over ESTs are based on three properties:

- Short sequence tags (10–20 bp) can contain enough information to uniquely identify a transcript, especially if the tag is obtained from a unique position in each of the transcripts. ESTs, on the other hand, are usually more than 350 bp long.
- Sequence tags can be concatenated to form long molecules. These molecules can subsequently be cloned and sequenced, allowing the serial processing of 25–50 transcripts in each sequencing run.
- The number of times that a particular tag is observed is a measure of the expression level of the transcript from which it is derived.

The SAGE tag is a nucleotide sequence of a defined length that is from a specific position in the transcript. The tag is usually directly adjacent to the 3′-most recognition site for a particular restriction enzyme in the cDNA of that transcript.

Basically, the SAGE method is as follows (Figure 6.4):

FIGURE 6.4. SAGE analysis. The first-strand synthesis is primed with a biotin-labeled oligo(dT) for later capture, and the second strand is synthesized. The cDNAs are bound to streptavidin columns and then digested with a 4-bp recognition restriction enzyme. The bound fragments are collected and are separated into 2 pools. Each pool is ligated to a different adaptor (1 and 2) and digested with the restriction enzyme whose recognition sequence is included in the adaptor releasing the tag from the remainder of the cDNA. The bound portion of the cDNA is removed. The 2 pools are combined, ligated together, amplified, and digested. The ditags are purified, concatenated, and cloned into a plasmid vector for sequencing. The sequences are deconvoluted to identify the cDNA tags and the tags clustered to find the number of times a particular tag is represented in the cDNA population. (Adapted from http://hg.wustl.edu/COGENE/INFO/sage-overview.html).


5'
AAAAAAAAAAAAA 3'
TTTTTT– B
First strand cDNA synthesis primed with biotin tabeled oligo(dT)
Second strand synthesis and digestion
B
Tag 1
Divide into two pools
Adaptor 1
Tag 1
B
Adaptor 2
Tag 2
B
Ligate
14–20 bp
14–20 bp
Adaptor 1
Tag 1
B
Adaptor 2
Tag 2
B
Digest
Adaptor 1
Tag 1
B
Adaptor 2
Tag 2
B
Combine pools 1 and 2 and ligate
Tag 1 Tag 2
PCR amplification and digestion
Tag 1 Tag 2
Concatenate
Tag 1 Tag 2
Tag 3 Tag 4
Tag 5 Tag 6
Tag 7 Tag 8
Sequence 25 – 50 tags per concatemer
CATGCCTGGATCTCGGTATGCTAGTCTGGTAGCTCGTGACCGTAGATCG
Tag 1
Tag 2
Tag 3
Tag 4
Ditag
Ditag

The mRNA population is converted to cDNA in which the first strand is primed with a biotin-labeled oligo(dT) and the second strand is synthesized as previously described. The double-stranded cDNA is digested with a restriction enzyme with a four-base recognition site that leaves a four-base overhang, for example, *Nla*III. The 3′ end of the digested cDNA is then captured on streptavidin-coated magnetic beads by using the biotin included with the oligo(dT). The pool of beads is split into two. Each pool is then ligated to a different linker molecule via the overhang introduced in the first digestion of the cDNA toward the 5′ end of the first strand of cDNA. These linkers each contain a recognition site for a type-2 restriction enzyme, such as *BsmF*1, that cuts at a specific distance past its recognition site, to allow the release of the linker-adapted SAGE tags. The tags from the two sample pools are repaired and blunt-end ligated to one another to form ditags (tags + linkers). PCR with primers to the two linkers is performed to amplify the heteromeric ditags. These amplified fragments are digested again with *Nla*III (if that was the initial enzyme used to digest the cDNAs) to release them from the linkers. The ditags are purified by PAGE and ligated to form long concatemers, which are size selected, gel purified, and cloned into a plasmid vector. Clones are then picked and sequenced. The length of the tags determines how unique that tag will be: the shorter the tag, the more the ambiguity, especially as the genome size increases. Thus 10-bp tags would not uniquely identify single sequences (because they occur approximately once every 10^6 bp by chance), whereas 20-bp or longer tags would be much more effective at unique identification (a 20-bp sequence would only occur approximately once every 10^{12} bp by chance).

The representation of the SAGE tags should accurately reflect the presence of those sequences in the transcript pool. The more frequently a particular tag appears, the more frequently that mRNA must be represented in the cDNA pool from which the tags were derived. Therefore, a comparison of the number of times a particular tag is found in the RNAs from different tissues or treatments gives a count that indicates the relative representation of that gene and its expression under the two treatments. Because each tag is short, the concatenation of the tags means that many genes are represented in each sequencing run. Therefore, although an abundant tag will be sequenced frequently, this method is less sequence intensive than an equivalent EST experiment (25–50 SAGE tags sequenced per run compared with a single EST per sequencing run). The additional depth of sequencing that is possible in SAGE experiments (up to 50-fold greater) will allow low-abundance messages to be identified by this method. The same low-abundance message would have been missed with an EST sequencing approach because the sequencing of the cDNA population would have been halted at an appropriate point with many fewer individual molecules sequenced.

MASSIVELY PARALLEL SIGNATURE SEQUENCING (MPSS™)

MPSS™ technology takes the SAGE approach a step further in the parallel processing of the sequence tags. Generation of sequence information from millions of DNA fragments is achieved by eliminating the individual sequencing reactions and the physical separation of DNA fragments. This technology is based on the ability to "clone" (ligate) cDNAs onto beads and then sequence in parallel hundreds of thousands of such beads (Figure 6.5).

As with all the expression profiling methods, a cDNA library is constructed from the appropriate tissue. In this case the oligo(dT) has a tail added that contains the restriction site for the enzyme to be used in the cloning of the cDNAs. The cDNAs are digested with a restriction enzyme and then ligated into a cloning vector that has a set of 1.67×10^7 different 32-mer oligonucleotide tags (Brenner et al., 2000). The cDNAs are amplified by using specific primers in the vector to expose the address tags, one of which also contains a fluorophore that is introduced at the end of the cDNA. The beads to which the cDNAs are attached each contain about a million copies of a single 32-mer antitag. The amplified cDNAs, each of which contains a unique tag, are then ligated to the mixture of beads so that each bead will only bind a single cDNA determined by the specific tag on that cDNA. Therefore, each bead will contain many copies of a single cDNA. The beads are then sorted by using a fluorescence-activated cell sorter to remove the beads that have not bound to a cDNA. This set of beads should contain a representative sample of the original cDNA library.

The beads that are complexed to a cDNA are then subjected to signature sequencing to determine a 16- to 20-nucleotide region from each of the cDNAs. This is achieved by immobilizing the beads as a monolayer in a flow cell. The sequences are read by iterative cycles that consist of ligating a short adaptor to the end of each cDNA and then using a restriction enzyme that digests remotely from its binding site and gives a 4-bp overhang at the cutting site. These sites are then specifically identified by using decoder oligonucleotides. Five rounds of such interrogation generate 20 bases of sequence for each cDNA. The ability to simultaneously sequence hundreds of thousands of beads means that even genes that are expressed at very low levels can be identified.

The characteristics of the MPSS™ system include:

- It sequences DNA molecules on as many as one million or more beads simultaneously.
- It eliminates the need for individual sequencing reactions and gels.
- It identifies each of the DNA molecules by a unique 16- to 20-base signature sequence.

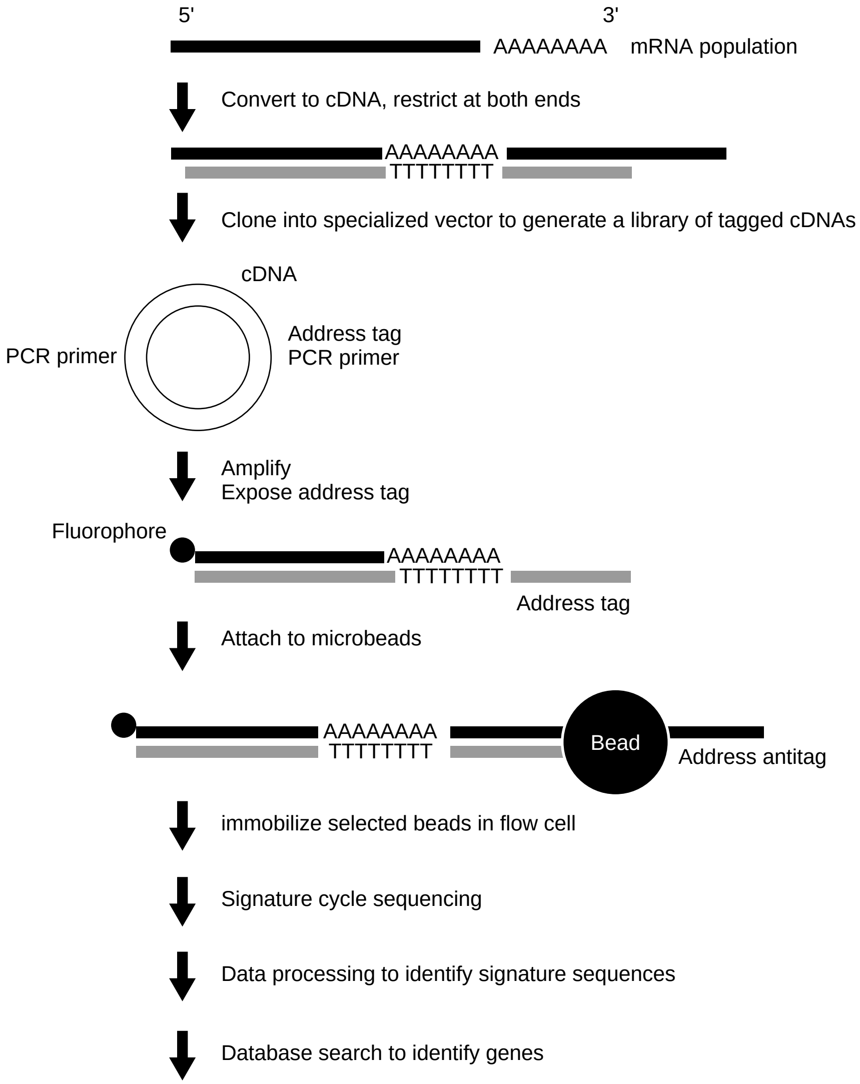


FIGURE 6.5. MPSS signature sequencing. The mRNA population is converted to cDNA with an oligo(dT) primer for the first strand that has a restriction site added to the 5′ end of the oligo. The cDNA is digested and cloned into a specialized vector. The cDNA is amplified and the address tag exposed and attached to a bead by the antitag. The cDNA signature sequence is obtained by cycle identification of 4 bp at a time. The signatures are used to search databases to identify the cDNAs. The signatures are clustered to identify the frequency at which any particular one is present in the cDNA population (Adapted with permission from Tyagi, 2000).

- It produces a comprehensive quantitative profile of gene expression in cells or tissues of interest.
- It has the potential to identify even the rarest expressed genes.

As with the data from SAGE and EST experiments, differential expression is detected by sequencing deeply into libraries and comparing the representation of the tags across the libraries. In this respect, the data analysis for MPSS is similar to that for SAGE. If a genome sequence or 3′ ESTs are also available, the origin of the tag and therefore the identity of the gene can be determined. These tags of 16–20 bp should be sufficiently unique to identify which particular gene in a family is being expressed, as discussed above with reference to SAGE technology.

THE ADVANTAGES AND DISADVANTAGES OF THE VARIOUS TECHNOLOGIES

As mentioned above, the techniques for expression profiling fall into two classes (comparative and digital). The microarray technique is an example of a comparative method, whereas ESTs, SAGE, and MPSS™ are all digital methods. Some of the differences between the various techniques are as follows:

- Microarrays can be used to monitor the expression of many thousands of genes simultaneously. However, they cannot give any information about the genes that are not present on the array (see comments concerning the design of the arrays above), so that a certain amount of preinformation is required. For any differences to be statistically significant sufficient replication of the biological samples, the replication of targets on each array and of the complete arrays must be included.
- EST sequencing gives a large amount of sequence information but also generates large quantities of redundant sequence, especially for those genes that are present in higher abundance. Although this redundant sequence can be reduced by using normalization or subtraction strategies, EST sequencing is not, in actuality, a viable method for looking at differential expression because of time and cost constraints.
- SAGE and MPSS strategies generate much sequence data and can be used for determining the actual proportions of various transcripts present. The SAGE technology has been used effectively for human transcription profiling studies, but relatively few data are currently available in plants. The drawback with MPSS™ is that it requires the use of proprietary Lynx technology and is expensive. However, a large amount of MPSS™ data for *Arabidopsis* is available at http://mpss.ucdavis.edu/java.html.

All of these technologies obviously require a selection of plant material as the basis of the investigations. This, in the end, may be one of the limiting factors in determining which of the techniques can be most appropriately applied. Where the biological material is the limiting factor it may be important to be able to use technologies that include amplification of the messenger RNA. For example, in looking for meiosis-specific transcripts from lily, it would certainly be useful to be able to the use the minimum amount of material possible. The paucity of other genomic resources may also be important in the choice of method. Again, taking the lily example, a lack of extensive knowledge of the genomic sequence (Chapter 4, Table 4.1) might make the interpretation of both MPSS and SAGE experiments difficult. However, in this same case, the decision as to what kind of microarray to use could also pose problems. One possible solution would be to use an *Arabidopsis* whole genome oligonucleotide array to see whether any "unexpected" transcripts were found and then to go back and look for their distribution and possible function in *Arabidopsis*.

An expression profile is a snapshot of the mRNA populations at a given moment in time. However, it gives no information about the stability of the RNA or about the translation rates of any given message. Therefore, whenever expression profiling is undertaken in a plant tissue it may be necessary to identify the protein components to maximize the information that can be gleaned from the expression profiles. The protein information will confirm the translation of the messages into proteins, as well as potentially identifying those proteins that are subsequently modified, whether for the purpose of their activation, inactivation, or degradation.

Proteomics

The advent of high-throughput technologies has facilitated a more holistic approach to the study of gene expression, allowing the coordinated characterization of many genes simultaneously, compared with previous studies that looked at genes or proteins individually (Anderson et al., 2000; Dutt and Lee, 2000; Lopez, 2000). Proteomics is the systematic analysis of "all" the individual proteins within a cell or tissue populations at a given time. These analyses should result in the characterization of all proteins simultaneously, as well as identifying their interactions. The characterizations should include the sequences and cellular localization as well as the identification of any posttranslational modifications and splice variants. The way in which proteins interact within the cell is additional information that will be essential for our understanding of cell functions. However, as with all the other high-throughput methodologies, data handling and analysis become critical issues (Patterson, 2001).

A proteomics experiment essentially consists of four steps:

- Sample preparation
- Protein separation
- Identification
- Functional analysis

High-throughput proteomics techniques used to characterize the protein complement include:

- Two-dimensional gel electrophoresis (2-DE)
- Image analysis
- Protein microsequencing
- Mass spectrometry

A sample flow of a proteomics experiment is shown in Figures 6.6 and 6.7. These figures illustrate the isolation of the proteins followed by their separation on 2-DE electrophoresis and the subsequent analysis of the individual separated proteins.

The choice of the proteins to be further characterized usually depends on the question being asked. For example, whenever comparisons are being made between tissues or treatments the 2-DE results are usually subject to image analysis and the differences in protein abundance of the various constituents identified (Jacobs et al., 2000). The protein spots can be characterized on the basis of peptide mass fingerprints by matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) and by partial amino acid sequencing. The information generated is sufficient for protein identification when coupled with database searches (Tabb et al., 2002; Lin et al., 2003). Initially, the protein identification was most effective in those cases in which large EST collections or genome sequences were available, so that the protein could be identified directly from the mass of the digest-produced peptides. However, the introduction of methods for the *de novo* peptide sequencing has meant that proteomics investigations can now be carried out effectively on proteins from species for which the nucleic acid sequence databases are insufficient. The use of both RNA and protein expression characterization has become an increasingly powerful combination for understanding the relationship between external perturbations and gene expression, as well as for identifying gene regulatory regions within the genome. Major technical challenges in plant proteomics will include the quantitative isolation of proteins for all compartments of the cell, the analysis of low-abundance proteins, the absolute quantification of expressed proteins, and the mapping of posttranslational modifications.

The disruption of the cell followed by the extraction of all of the proteins only allows the characterization of the individual protein molecules. Frequently, however, the information that is vitally important is how particular

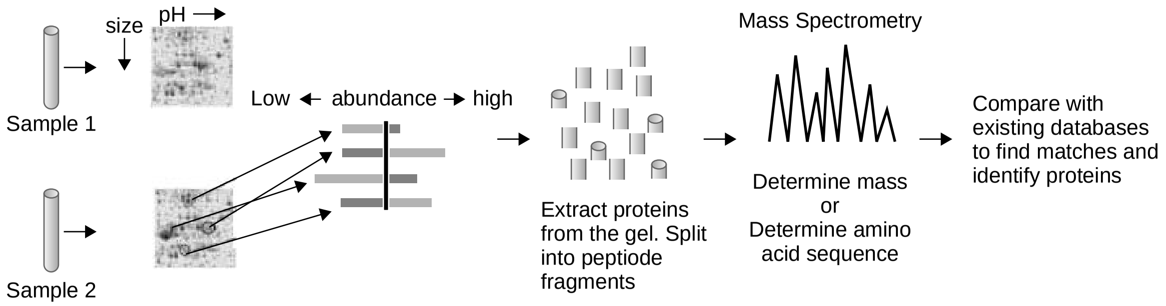


FIGURE 6.6. Proteomics experimental flow. Isolation of proteins followed by their separation by 2-D gel electrophoresis. The pattern of the proteins is compared, and those of interest are excised, fragmented, and separated by mass spectrometry. The amino acid composition or sequence is determined (depending on the particular MS technology applied), and the databases are searched to identify the protein. The actual amino acid sequence is important for those species for which there is little genomic or cDNA sequence. ▢ Protein abundance from sample 1; ▣ Protein abundance from sample 2.

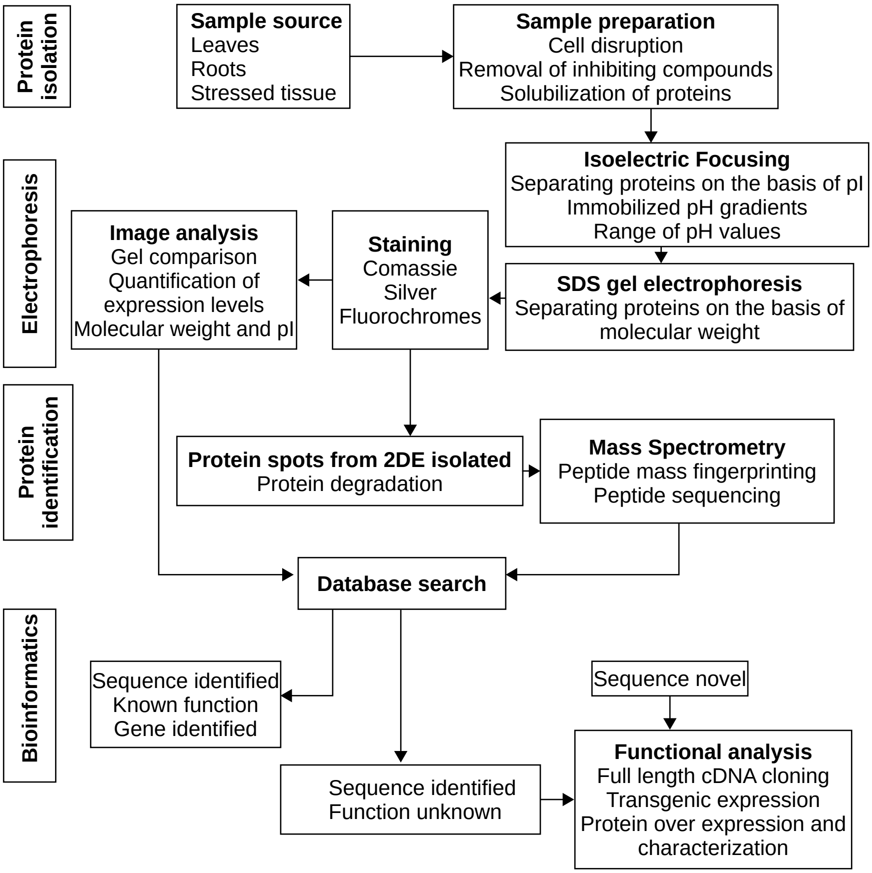


FIGURE 6.7. A schematic overview of the interconnections in proteomics (reproduced from Jacobs, D. I., R. van der Heijden, and R. Verpoorte (2000) Proteomics in plant biotechnology and secondary metabolism research. *Phytochem. Anal.* **11**, 277–287).

proteins interact. Therefore, an important aspect of proteomics is the isolation of protein complexes and their subsequent characterization to understand the *in vivo* interactions.

PROTEIN ISOLATION

The overall performance of proteome analysis depends on the initial steps, namely, the isolation of the proteins and the separation tool applied to these isolated proteins. Plants are differentiated organisms, and many of the spe-

cialized functions are carried out in particular tissues or organs. Each cell type is likely to contain a specific set of proteins, and so the choice of plant material will determine the particular proteome to be analyzed (Jacobs et al., 2000). For example, the flavor and aroma compounds found in basil are synthesized on the leaf surface in specialized peltate glandular trichomes (glands), and so this would be a source for identification of the proteins involved in the synthesis of various secondary compounds. For the characterization of all of the individual proteins the ideal sample preparation would result in all noncovalently bound proteins as a solution of individual polypeptides, with the coincident removal of any interfering compounds, of which plants have many.

Unfortunately, plants are not ideal as a source of proteins for isolation with 2-DE. The protein content is relatively low, and a large part of the plant cell is occupied by the vacuole, which is filled with compounds that have a negative effect on protein extraction and chromatography. A number of methods for protein extraction have been described (Jacobs et al., 2000) to minimize the effects of inhibitory or interfering compounds. Finally, after extraction, the proteins must be solubilized in a buffer that is appropriate for the subsequent separations. There is not a single method that can be used in every instance, so that the choice of sample preparation is dependent on the aim of the investigation. The most complete overview of the proteins from a single sample is likely to require a combination of different methods, the results of which will subsequently need to be combined.

PROTEIN SEPARATION

Protein separation and identification can occur either through separation by 2-DE or, more recently, by the application of liquid chromatography methods.

The 2-DE methodologies usually use a first-dimensional separation of the proteins in a pH gradient, whereas in the second dimension proteins are separated according to their molecular weight. The introduction of immobilized pH gradients (IPG) for the first-dimension isoelectric focusing (IEF) has added flexibility to the magnitude of the pH range in the first-dimension separations and so has resulted in an increase in the resolving power of the second-dimension polyacrylamide gel electrophoresis (PAGE). After the gel separation the proteins are visualized with a variety of possible staining methods, including Coomassie and silver staining. One of the considerations in choosing the staining protocol is the linear dynamic range of the stain, because the range of protein concentrations in plant extracts can be greater than 10^6 (Corthals et al., 2000). When comparing the proteome under different conditions, both the qualitative and quantitative variations in individual proteins will be important. Silver staining is more sensitive than Coomassie

but has a much more restricted dynamic range. A number of fluorescent dyes have been developed that provide the sensitivity of silver staining but with a larger dynamic range. The use of fluorescent dyes that can be coupled to different samples, followed by mixing of the samples and separation on the same gel, will facilitate the direct visualization of the differences between the two samples. After staining the gel is subject to image analysis, from which a quantitative determination of the protein spots can be made. Molecular weight and pI can be calculated for each of the proteins.

PROTEIN IDENTIFICATION

The application of mass spectrometry (MS) for protein identification and amino acid sequencing continues to improve (Lin et al., 2002). For proteins separated on 2-D gels, the protein spot can be directly excised and subjected to MS techniques for identification. The molecular weights of the peptide fragments, on their own, are insufficient for protein identification. However, when molecular weight is allied with peptide fingerprinting and sequence information, protein identification is possible. Two forms of protein identification are available:

- An unknown protein is enzymatically digested, and the masses of peptides produced are determined (mass mapping). These fragments are used to search the databases to identify a protein that would produce fragments that match the data, and so the protein can be identified. Data from cDNAs and genomic sequences can be used in the searches.
- Tandem MS of peptide fragments can reveal their actual amino acid sequences. These peptide sequences can again be used in database searches to identify the structure of the complete protein.

Both of these methods will handle proteins separated on gels. However, the development of tandem MS, which can also be used to handle protein mixtures, has increased the possible options in proteomics. "Shotgun proteomics" is a term coined to describe the process of characterizing a complex protein mixture with the identification of the proteins originally found in the sample (Tabb et al., 2002).

PROTEIN-PROTEIN INTERACTIONS

The most common extraction and separation techniques are designed to eliminate protein-protein interactions and therefore are only useful for detecting and characterizing the individual peptides. However, one possible

way to assign a function to a protein is to identify the other proteins with which it interacts. The discovery and characterization of protein-protein interactions has become easier with improvements to the yeast two-hybrid system, protein tagging, and advances in mass spectrometry.

Yeast Two-Hybrid Systems

The yeast two-hybrid system is a powerful tool for discovering protein-protein interactions *in vivo* and involves screening "prey" proteins encoded by cDNA libraries for interactions with a particular "bait" protein (Allen et al., 1995; Chien et al., 1991). This system involves the cloning of cDNAs into vectors that fuse the open reading frame with either a DNA binding domain (bait construct) or an activation domain (prey construct). The function of these two constructs is that when they are brought together they can activate a reporter gene. The two domains can only be placed in the appropriate structure if the bait and prey proteins interact with each other. The two separate libraries are transformed into yeast, and then the two different plasmids (the "bait" and "prey") are introduced into a single yeast cell. The two plasmids are combined in a single cell by overlaying replica grids of the sets of bait and prey colonies. The cells mate, and the resultant cells contain both bait and prey plasmids (Figure 6.8). If the bait and prey proteins interact, the DNA binding and transcription activation domains are brought into close proximity and activate transcription of the reporter gene.

By applying the yeast two-hybrid system on a genome-wide basis all the pair-wise interactions can be identified, as, for example, if all the *Arabidopsis* open reading frames were cloned into a two-hybrid activation domain (AD) vector, thereby expressing 25,000 His-AD fusions. All yeast cells expressing these fusions could then be mated to yeast strains that express a fusion of the His-DNA-binding domain (DBD) and the particular *Arabidopsis* open reading frames under investigation. These experiments only give the interactions between two proteins, but they can obviously be extended in a linear fashion by using the newly identified interacting proteins as "baits."

Protein Tags and Transgenics

An alternative to using the yeast two-hybrid system is to tag a protein to enable the isolation of the *in vivo* complexes intact (Honey et al., 2000). Therefore, the combination of tagging a full-length cDNA, with an extension that will enable its protein product to be isolated, and the transfer of the tagged gene back into the organism could result in the purification of intact complexes of proteins (Figure 6.9). The affinity-tagged protein must maintain protein function in order to undergo the correct functional interactions. In addition, it is important that the introduced tagged copy of the gene is

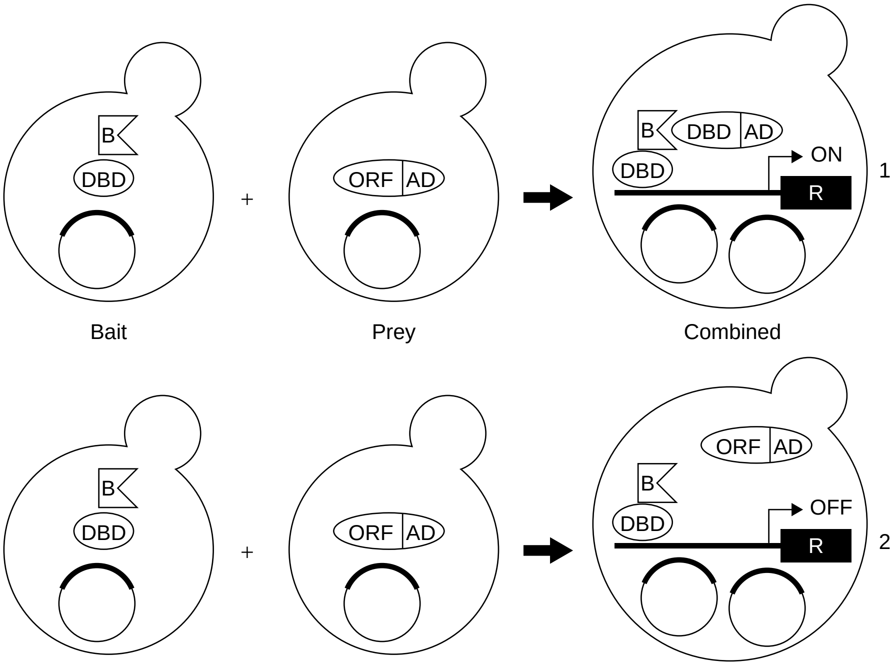


FIGURE 6.8. The yeast 2-hybrid system. Full-length cDNAs are cloned into one of two constructs, the bait or prey. The bait construct fuses the introduced protein (B) to a DNA binding domain (DBD). The prey construct fuses the introduced protein (ORF) to an activation domain (AD). The bait is chosen, and the line is mated to the whole library of prey proteins to generate cells containing both bait and prey constructs. (1) The bait and prey proteins interact, placing the DNA binding domain and the activation domain together so that they bind to and activate the reporter gene. (2) The bait and prey proteins do not interact so that, although the DNA binding domain is still bound, the activation domain is not present so the reporter gene is not turned on (Adapted with permission from http://depts.washington.edu/sfields/yp_project/YPLM.html).

expressed in the same location as the endogenous protein. An alteration in the pattern of expression would allow possible artifactual interactions to be characterized. However, at a first approximation, any information about potential protein complexes and the molecules involved will be important in understanding the function of that protein *in vivo*. The ideal situation would be to use the tagged construct to replace the endogenous gene so that the tagged protein would be expressed at the correct stage and tissue. However, homologous recombination in plants is not sufficiently efficient at present for this to be accomplished, so the introduced copy must be placed under the control of other promoters. The overexpression of the tagged

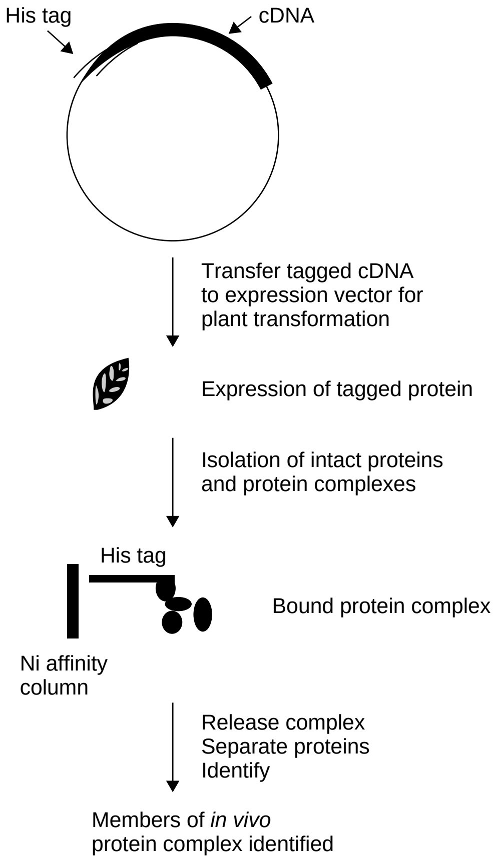


FIGURE 6.9. Isolation and characterization of *in vivo* protein complexes. A full-length cDNA is inserted in frame into a vector that will add a His tag to the protein. This construct is transformed into a plant under the control of an appropriate promoter. The transgenic plant is grown and the tagged protein extracted under conditions in which the protein complexes remain intact. The complex is isolated by binding the complex with the tag (in this example the His tag bound to a nickel cartridge), released and the components dissociated, and characterized by 2-DE and MS to identify all the protein components of the complex.

protein or its presence in tissues where it is not normally found introduces the possibility of generating artifactual interactions.

Summary

The development of high-throughput tools to look at the transcriptome and proteome has made a huge impact on our approach to understanding plant form and function. The ability to simultaneously measure the level of tran-

scription of all the genes will increase our understanding of the coordination of gene expression at the transcription level. Thus the high-throughput methodology for expression profiling is a way to focus on the interesting genes while still being able to have a comprehensive view of gene expression. The more detailed characterization of the fluctuations of each of these interesting genes will still need to be done. One method for these extended studies and for confirmation of the expression profiles is real-time PCR. This will be useful for the characterization of the detailed dynamics of any changes in expression. The coupling of the expression profiling data with the characterization of the protein components of the cell facilitates the understanding of the transfer of information from DNA to RNA to proteins. An additional benefit to these studies will be the detection of those RNAs that are not immediately, or ever, translated into proteins. The detailed protein studies will also shed light on the role of protein modification in protein function. High-throughput metabolomic studies are still in their infancy, but the addition of these data will help our understanding of the flow of information through all of the control points in the cell from genomic sequence to final product.

REFERENCES

Allen, J. B., M. W. Wallberg, M. C. Edwards, and S. J. Elledge (1995) Finding prospective partners in the library: the yeast two-hybrid system and phage display find a match. *TIBS* **20**, 511–516.

Anderson, N. L., A. D. Matheson, and S. Steiner (2000) Proteomics: applications in basic and applied biology. *Curr. Opin. Biotechnol.* **11**, 408–412.

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao. and K. Corcoran (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18** 630–634.

Chien, C. T., P. L . Bartel, R. Sternglanz. and S. Fields (1991) The two-hybrid system: A method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. USA* **88**, 9578–9582.

Churchill, G. A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat. Genet. Suppl.* **32**, 490–495.

Corthals, L., V. C. Wasinger, D. F. Hochstrasser, and J.-C. Sanchez (2000) The dynamic range of protein expression: A challenge for proteomic research. *Electrophoresis* **21**, 1104–1115.

Dutt, M. J., and K. H. Lee (2000) Proteomic analysis. *Curr. Opin. Biotechnol.* **11**,176–179.

Grabowski, P. (2002) Alternative splicing in parallel. *Nat. Biotechnol.* **20**, 346–347.

Honey, S., B. L. Schneider, D. M. Schieltz, J. R. Yates, and B. Futcher (2001) A novel multiple affinity purification tag and its use in identification of proteins associated with a cyclin-CDK complex. *Nucleic Acids Res.* **29**, 24.

Jacobs, D. I., R. van der Heijden, and R. Verpoorte (2000) Proteomics in plant biotechnology and secondary metabolism research. *Phytochem. Anal.* **11**, 277–287.

Kane, M. D., T. A. Jatkoe, C. R. Stump, J. Lu, J. D. Thomas, and S. J. Madore (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res*. **15**, 4552–4557.

Lin, D., D. L. Tabb, and J. R. Yates (2003) Large-scale protein identification using mass spectrometry. *BBA Proteins Proteomics* **1646**, 1–10.

Lipshutz, R.J., S. P.A. Fodor, T. R. Gingeras, and D. J. Lockhart 1999 High density synthetic oligonucleotide arrays. *Nat. Genet. Suppl.* **21**, 20–24.

Lockhart, D. J., and E. A. Winzeler (2000) Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836.

Lopez, M. F (2000) Better approaches to finding the needle in a haystack: Optimizing proteome analysis through automation *Electrophoresis* **21**, 1082–1093.

Madden, S. L., C. J. Wang, and G. Landes (2000) Serial analysis of gene expression: from gene discovery to target identification. *Drug Discovery Today* **5**, 415–425.

Modrek, B., and C. Lee (2002) A genomic view of alternative splicing. *Nat. Genet.* **30**, 13–19.

Patterson, S. D. (2001) Data analysis—the Achilles heel of proteomics. *Nat. Biotechnol.* **21**, 221–222,

Powell, J. (1998) Enhanced concatemer cloning - a modification to the SAGE (Serial Analysis of Gene Expression) technique. *Nucleic Acids Res.* **26**, 3445–3446.

Soellick, T. R., and J. F. Uhrig (2001) Development of an optimized interaction-mating protocol for large-scale yeast two-hybrid analysis. *Genome Biol.* 2001, **2**, research 0052.1–0052.7.

Tabb D. L., W. H. McDonald, and J. R. Yates (2002) DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1**, 21–26.

Tyagi, S. (2000) Taking a census of mRNA populations with microbeads. *Nat. Biotechnol.* **18**, 597–598.

Velculescu, V.E., L. Zhang, W. Zhou, B. Vogelstein, and K.W. Kinzler (1995) Serial analysis of gene expression. *Science* **270**, 484–487.

CHAPTER

7

INTERACTIONS WITH THE EXTERNAL ENVIRONMENT

OVERVIEW

In nature, plants encounter a wide range of environmental conditions from which they cannot escape. Therefore, they have a number of mechanisms with which to counter these slings and arrows of outrageous environment. The responses must be available at all stages of development. The stresses can be biotic in the form of pests and diseases or abiotic, with extremes of temperature or moisture being particularly common. As well as being an interesting topic in itself, understanding how sessile organisms can cope with a fluctuating environment has a practical importance in agriculture. Salt stress is a problem in many parts of the world, especially on irrigated land. Drought stress is a perennial problem, especially in those parts of the world where food security is in the balance. Pests and diseases are major factors of reduced yield and the need for chemical spraying resulting in environmental pollution. Therefore, an understanding of the processes by which plants have adapted to overcome these limitations, and the detailed characterization of the pathways involved, may enlighten the search for durable disease- and stress-tolerant plants in agriculture.

Studies on stress-related responses must include a whole range of considerations from plant morphology through to gene expression to draw meaningful conclusions about the root causes of adaptations. They will need to be performed in species in which the molecular, physiological, and metabolic data are integrated with phenotypic descriptions. This will allow for the creation of integrated models that can describe how individual genes and metabolites work together in a coordinated response to various stresses.

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

The availability of microarrays has changed the way in which gene expression in response to various stresses can be assayed. The unraveling of the complex interactions between plants and microbes will be made easier with the addition of the information from microbial genomics. The relative contributions of host and pathogen in determining whether or not an interaction will result in disease or elicit a resistance response are starting to be identified. The addition of proteomics allows the correlation of the expression data at the functional level and also permits the importance of post-translational modifications of proteins in any responses to be identified. All of this new information must be integrated to understand both the common and unique responses in the wide variety of interactions between the plant and its environment.

The interactions of plants with their environment is an extensive and complex subject that cannot be covered in detail here. An inspection of the projects funded through the Plant Genome Research Program at the National Science Foundation (http://plantgenome.sdsc.edu/) and the USDA Cooperative State Research Education and Extension Service (CSREES) program gives a good indication of the scope of the subject. Provided here is a glimpse of how genomics and proteomics are impacting this area of study.

Biotic Interactions

Plants come into contact with a wide range of other organisms in their lifetime. These organisms can be beneficial, harmful, or neutral, and the plant must be able to respond appropriately to each type. Thus plants have developed sophisticated mechanisms through which they respond to these organisms. For example, how does the plant distinguish between a pathogenic invasion and the establishment of a symbiotic relationship, and then respond appropriately by excluding or destroying the former while guiding the development of the latter? High-throughput genomics approaches have made the global study of these responses possible. The added information becoming available from the sequencing of the pathogen and symbiont genomes will contribute to the development of a complete understanding of the genes involved in the interactions. It is clear that in the interactions between microbes and plants the same mechanisms are often used irrespective of whether the microbe is beneficial or pathogenic. However, the integration, timing and combinations of these mechanisms may result in the different outcomes elicited by specific organisms.

Disease Resistance

The original gene-for-gene disease resistance interaction in plants was defined by Flor (1971) for the interaction between flax and its pathogen flax

rust. In this interaction the host and pathogen must have a pair of specific alleles, one in each organism, for resistance to occur. The resistance gene (R) in the plant and the corresponding avirulence gene (Avr) in the pathogen are the basis of resistance, generally eliciting the hypersensitive response (Table 7.1). The hypersensitive response occurs after the recognition of the pathogen by the plant with the induction of a cell death cascade in the vicinity of the infection.

Disease resistance (R) genes and their homologs are among the most prevalent genes in plant genomes (Meyers et al., 1999). These resistance genes have been the subject of intense scrutiny over the past decade, and many have been isolated. The isolated genes generally fall into five classes (Hammond and Jones, 1997):

- The nucleotide binding (NBS)-leucine-rich repeat (LRR) genes
- Detoxifying enzymes, e.g., the Hm1 gene in maize
- Intracellular serine/threonine protein kinases, e.g., the Pto gene in tomato
- Extracellular LRR proteins with a single membrane spanning region and short cytoplasmic carboxyl terminus, e.g., the cf9 gene in tomato
- Extracellular LRR proteins with a single membrane spanning region and cytoplasmic kinase domain, e.g., the Xa21 gene in rice

The major class of these resistance genes is the NBS-LRR group, and this group can be further subdivided into two subclasses (Jones, 2001):

- Those that carry amino-terminal homology to the Toll and Interleukin-1 receptor (TIR) genes (the TIR-NBS-LRR family), e.g., the L_6 gene in flax

TABLE 7.1. THE GENE-FOR-GENE INTERACTION BETWEEN PLANT RESISTANCE GENES AND PATHOGEN AVIRULENCE GENES.

	Plant Genes	
Pathogen genes	R1	R2
Avr1	Hypersensitive response Cell death	No resistance, disease progression
Avr2	No resistance, disease progression	Hypersensitive response Cell death

The presence of an interacting pair of genes results in the activation of the hypersensitive response, leading to cell death. Other pairs of resistance and avirulence genes do not activate the hypersensitive response and so result in disease progression.

- Those that carry a putative coiled coil (CC) at their amino terminus (the CC-NBS-LRR family), e.g., the RPS2 gene in *Arabidopsis*.

The structure of the R genes in these classes is shown in Figure 7.1.

These R and Avr genes are shaped by a continuous evolutionary battle. The plant R genes are believed to recognize the products of the pathogen Avr genes and to activate the defense program (the hypersensitive response) that prevents the pathogen from gaining a foothold. Because the Avr genes prevent the colonization of the plant by the pathogen, selection should have eliminated them unless they perform another essential function in the pathogen, perhaps ensuring virulence in nonresistant plants. There is

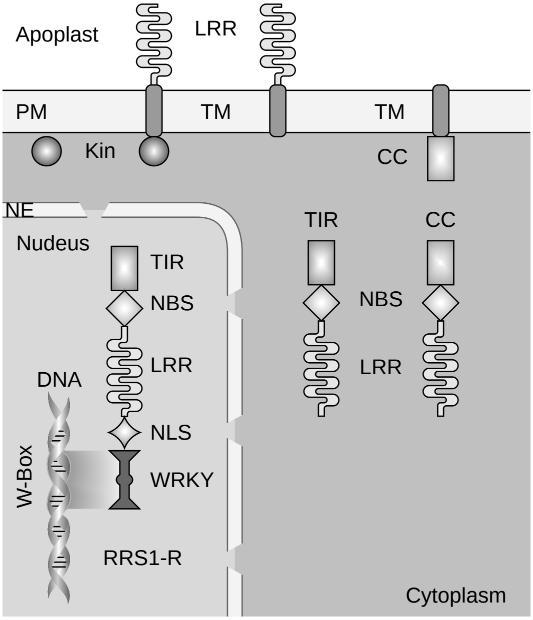


FIGURE 7.1. Modular composition and predicted location of R protein classes. NBS-LRR proteins are predicted to encode cytoplasmic receptor molecules. RRS1-R represents a novel NBS-LRR type that encodes a C-terminal NLS and a WRKY domain. Presence of the WRKY DNA-binding domain suggests that RRS1-R activates genes that are under the transcriptional control of W-box-containing promoters. Abbreviations: CC, coiled-coil domain; Kin, kinase; LRR, leucine-rich repeat domain; NBS, nucleotide-binding site; NE, nuclear envelope; NLS, nuclear localization signal; PM, plasma membrane; TIR, Toll/interleukin-1-receptor; TM, transmembrane domain. (Reprinted from *Trends Plant Sci.* **7**, Lahaye, The *Arabidopsis* RRS1-R disease resistance gene—uncovering the plant's nucleus as the new battlefield of plant defense, 425–427, Copyright 2002, with permission from Elsevier.)

evidence of such a role for virus-encoded proteins that are recognized by host R genes (Nimchuk et al., 2001).

Recognition is considered to be the initial important event in the response of plants to microbes. Many of the genes in the pathogen that are responsible for both the invasion and the activation of defense responses are thought to encode secreted proteins from the pathogen. One of the mechanisms used by pathogens to deliver these proteins is a specialized secretion system called the type III secretion system (TTSS), which delivers the bacterial proteins directly into the host cell (Lugtenberg et al., 2002). This system is widespread among both animal and plant pathogens and, as noted below, may also play a role in symbiosis (Staskawicz and Parniske, 2001). Therefore, the genomic sequences that are available from some pathogens can be used with informatic tools to identify which of the open reading frames are likely to be secreted proteins based on the structure of their leader sequences. These proteins are primary targets for investigating the roles and interactions between R and Avr genes. The identification of these relationships and the roles of secreted proteins in the development of disease will be vital for the development of new points of intervention in disease processes in crop plants.

An understanding of the signal transduction networks by which the R genes control the activation of defensive responses is of considerable interest because the R genes are generally quite effective in preventing disease. The use of *Arabidopsis* and the insertion mutants that are available has thrown considerable light on these networks and the genes involved (Glazebrook, 2001). These data indicate that there are at least three distinguishable mechanisms through which gene-for-gene resistance can be mediated. Some of the proposed interacting pathways are shown in Figure 7.2.

Two of these pathways involve intermediate genes that have been identified, whereas the third does not yet have any identified required intermediates between the R gene and the resistance reaction. The genes involved in mediating the resistance responses following the interaction of the R and Avr gene products have the following properties, as shown by mutant studies:

- NDR1—Required for non-race-specific resistance to bacterial and fungal pathogens. It also mediates the systemic acquired resistance (SAR) response. This gene can operate in both linear and parallel signaling events, depending on the R gene function triggered (Tornero et al., 2002). The *Arabidopsis* genome contains 28 genes with sequence homology to the *Arabidopsis* NDR1 gene. Expression analysis of eight of these genes identified two that show pathogen-dependent mRNA accumulation (Varet et al., 2002). One of these two genes was also expressed during infection with an avirulent oomycete, *Peronospora parasitica*.

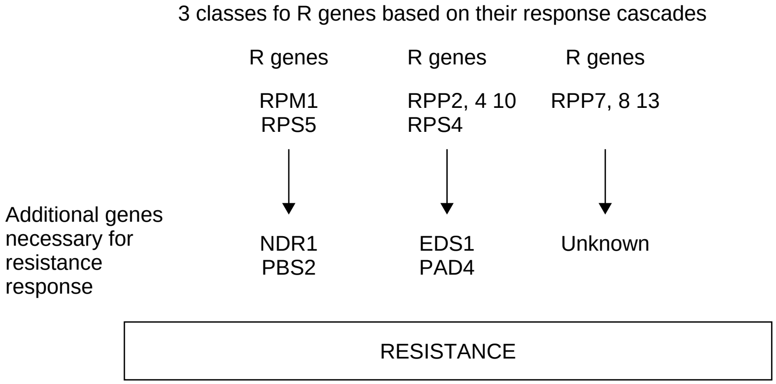


FIGURE 7.2. Possible positions of some of the *Arabidopsis* genes in signal transduction networks that control the effect of resistance genes. The box containing the word "resistance" represents the final expression of resistance. Three R-gene-dependent pathways are shown, one that requires *NDR1* and *PBS2*, a second that requires *EDS1* and *PAD4*, and a third for which the required genes have not been reported. (Reprinted from *Curr. Opin. Plant Biol.* **4**, Glazebrook, Genes controlling the expression of defense responses in *Arabidopsis*—2001 status, 301–308, Copyright 2001, with permission from Elsevier.)

- PBS2—A mutant in this gene has reduced RPS5- and RPM1-mediated resistance. It is potentially involved in transduction of R gene-mediated disease resistance.
- PAD4—Encodes a lipaselike gene that is also important for salicylic acid signaling. PAD4 is also required for runaway cell death in the lsd1 mutant. Importantly, this novel function of PAD4 is operative when runaway cell death in lsd1 is initiated through an R gene that does not require PAD4 for disease resistance (Rusterucci et al., 2001).
- EDS1—This is a component of R gene-mediated disease resistance in *Arabidopsis thaliana* with homology to eukaryotic lipases. EDS1 is essential for disease resistance conferred by the structural subset of resistance (R) proteins, the TIR-NBS-LRR proteins, but is not required by the CC-NBS-LRR proteins (Peart et al., 2002). EDS1 is also required for runaway cell death in the lsd1 mutant. Importantly, this novel function of EDS1 is operative when runaway cell death in lsd1 is initiated through an R gene that does not require EDS1 for disease resistance as is also the case for PAD4 (Rusterucci et al., 2001).

It is clear that each of these genes can have effects in various other resistance pathways and not just in interactions with the R genes.

In addition to the direct pathways from the R gene to resistance, there are also the salicylic acid-mediated responses and jasmonic acid/ethylene-

mediated responses (Glazebrook, 2001). The salicylic acid-dependent signaling response is important for SAR that is activated throughout a plant in response to particular types of infection. Some of the same genes, for example, the PAD4 gene, that are involved in resistance mediated by the gene-for-gene pathway are also involved in the salicylic acid signaling pathway. The SAR response and the signaling response to jasmonic acid/ethylene appear to be much more complex than the direct pathway between the R gene and the expression of resistance. As more mutants involved in these pathways are identified (and multiple mutant analyses performed) the interrelationships and dependence of each of the points in the pathways with respect to each other are likely to become clearer.

Genome-wide expression profiling that includes mutants affecting disease resistance responses will result in the identification of the genes important in the regulation and expression of disease responses. Disease resistance pathways have many common elements that overlap with the responses induced by other stresses. Therefore, there exists a network of genes that are activated in response to a wide range of stimuli but do not have any essential role in the development of resistance *per se*. The development of these expression profiles for a wide range of stimuli will therefore help in identifying those genes that are integral and essential for the expression of specific resistances.

As has been indicated in other chapters, the fact that a gene is observed to have differential expression under various conditions does not necessarily lead to an understanding of its role in the phenomenon under investigation. Rather, the level of product associated with that transcript and the proteins with which that product is intimately associated in the cell to mediate the responses are essential, but currently missing, pieces of information. The LRR domains of the R proteins are thought to act as the determinants of specificity and ligand binding. For example, in the yeast two-hybrid system an interaction between the LRR-like domain of the rice resistance protein Pita and the Avr-Pita protein was demonstrated (Nimchuck et al., 2001). However in the complex signaling pathways, especially those involved in SAR and the jasmonic acid/ethylene signaling responses, the extent and number of proteins that form complexes with the resistance proteins has yet to be determined. The characterization of the extent of these interactions should go a long way toward providing an understanding of the basic mechanisms by which the R genes act, and how pathogens can bypass the defenses, and so direct attention to possible positions in these pathways for interventions that can lead to new methods of disease resistance for crop plants.

The unraveling of the importance and contributions all of the players will use the whole suite of current genomics methodology. Thus:

- Expression profiling will identify those genes that are modulated in response to disease challenge in both the host and the pathogen.

- Proteomics techniques, especially those using transgenics, will allow the isolation of the in vivo protein complexes and the identity of the involved components.
- Insertional mutagenesis, and other mutagenesis studies, will identify the involvement of each of these genes.

Considering the apparent redundancy of the pathways through which the resistance is developed, the unraveling of these pathways by mutant analyses is likely to involve various combinations of multiple mutants in a single line. These experimental determinations will be necessary to fully understand the roles of each of the players in mediating the wide range of resistance responses.

To possibly use resistance genes across wide species barriers, the resistance phenomena will need to be studied across both a wide range of plant species and a wide range of pathogen types. Such studies will necessitate the characterization of the interacting proteins and the regions of the proteins involved in such interactions. The introduction of resistance genes into heterologous systems has only conferred resistance when the transfers have been among species that are closely related. Therefore, understanding the reasons for this may elucidate any specific evolutionary constraints that have been imposed across the plant kingdom. Natural selection has been acting on the R gene loci over much longer periods than those during which plant breeders have been recruiting the various forms of these genes for crop improvement (Jones, 2001). The effect on fitness of changing the R gene profile, particularly with respect to "stacking" R genes to provide more durable resistances, also must be considered. Because R genes are thought to be always expressed and to function as alert first responders, the stacking of many resistance genes into a single genotype may not provide the desired durable resistance, while still leading to a reduction in yield in the absence of the pathogen.

Response to Symbiosis

The interaction of plant roots with the soil environment is a much less well-developed area of research and knowledge compared with the study of above-ground interactions. The roots mine the soil for nutrients, and the nutrient uptake is affected by symbiotic interactions with arbuscular mycorrhizal fungi, which supply plants with phosphate and other nutrients. In legumes the development of the nitrogen-fixing nodules is a highly specific interaction between the appropriate bacteria and the response of the legume host plant. This interaction allows for the invasion of the plant root, with the cooperation of the plant, in the form of the development of an infection thread and the formation of the nodule, so that the bacteria are provided with the appropriate environment in which they can fix nitrogen. So how is

this invasion controlled and how does the plant host differentiate it from the reaction with pathogenic bacteria? Similarly, how are the fungal hypha accommodated during colonization by mycorrhizal fungi without causing a plant defense reaction?

Because *Arabidopsis* is a poor host for mycorrhiza and does not undergo nodulation by *Rhizobium*, it is not a particularly suitable model system for the study of symbiosis. As was the case for disease resistance, the symbiotic relationship must be considered from two points of view, namely, those of the host plant and the symbiont.

The lack of a genomic sequence for any of the legumes has certainly slowed the pace at which the genes that are responsible for symbiosis have been isolated. Despite this, numerous plant mutants that affect the legume-*Rhizobium* symbiosis in particular have been isolated, and these mutants define more than 50 symbiotic loci that are indispensable for either or both the formation and functioning of a root nodule and the formation of the arbuscular mycorrhiza symbiosis (Staskawitz and Parniske, 2001). Specifically, the mutants available in the model legumes *Lotus japonicus* and *Medicago truncatula* can be divided into four classes on the basis of their phenotypes (Stougaard, 2001):

- Nonnodulating—These mutants are expected to be impaired in the early events necessary for the establishment of the infection.
- Ineffective nodulating mutants—These mutants can participate in the early events associated with the development of nodules, but the process cannot be completed so that either nodule organogenesis or nodule function cannot be completed.
- Mutants with increased or decreased nodule numbers—Supernodulation, the increase in nodule numbers, could result from mutations in genes that regulate the nodule number. A reduced number of nodules may be caused either by mutations that regulate nodule number, as with supernodulation mutants, or by mutations that impair nodule formation but do not prevent it from occasionally developing to completion.
- Mutants with delayed nodulation.

Missing from the mutant types in the model legumes are the spontaneous nodulation and "tumorlike" nodules that have been described in alfalfa (Stougaard, 2001).

In many cases the events occurring on the microbial side of the symbiotic relationships have been better characterized because of the relative ease of working with the symbiont compared with the root system. The sequencing of the symbiotic chromosome in broad host-range *Rhizobium* NGR 234 has led to the characterization of the type III secretion systems that had previously been thought to be unique to pathogenic bacteria (Marie et al., 2001).

Again, genomic approaches will have a major impact on the separation of the different processes that determine whether or not an interaction will progress down the path of symbiotic coexistence or pathogenic destruction.

The processes of infection and nodule development have been well documented, and therefore the isolation of mutants involved in nodulation has been possible. The interaction between the plant and mycorrhiza has been more difficult to detail because the process is not as amenable to study. However, a number of genetically defined loci that control some of the steps in both the nodulation and mycorrhizal interactions have been identified. A schematic view of the early symbiotic events in both of these processes is provided by Figure 7.3.

The seemingly complex interactions that control pathogenesis and symbiosis, and how the two interactions differ from one another, is amenable to being unraveled with the genomics and proteomics tools currently available. The development of custom microarrays with ESTs, and putative genes identified from complete genomic sequences, from the various stages of infection (both symbiotic and pathogenic) including representatives from both of the players in the interaction will facilitate the identification of the important genes involved in the control of the interaction. The use of yeast two-hybrid systems to determine the interactions between plant and microbial proteins at various stages of the process will also identify the important reactions necessary for resistance, susceptibility, or symbiosis. The development

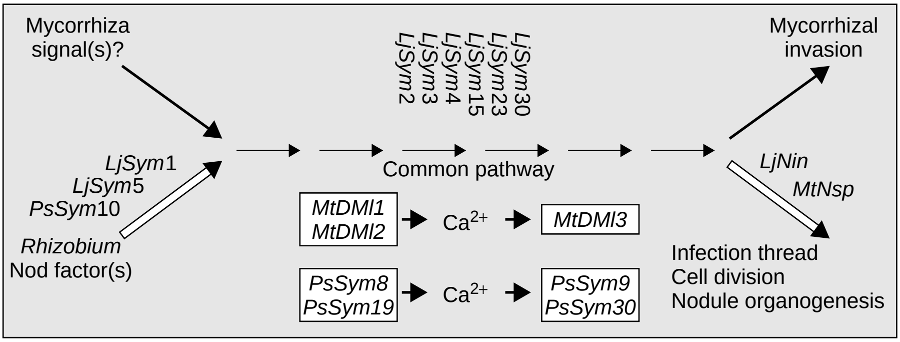


FIGURE 7.3. Analysis of *Lotus*, *Medicago*, and pea mutants for rhizobial and mycorrhizal interaction defines common steps in the endosymbiotic pathway. The allocation of symbiotic loci to various parts of the pathway is tentative and is based on phenotype, rather than on genetic tests of epistasis. The order of appearance does not therefore indicate the precise order of function. Common pathway genes *MtDMI1*, *MtDMI2*, *MtDMI3*, *PsSym8*, *PsSym19*, *PsSym9*, and *PsSym30* were ordered according to calcium spiking. Except for *Ljnin*, none of the genes have been cloned. (Reprinted from *Curr. Opin. Plant Biol.* **4**, Stougaard, Genetics and genomics of root symbiosis, 328–335, Copyright 2001, with permission from Elsevier.)

of metabolomics will be important in defining the roles of signaling molecules in the initiation and progression of all of these interactions. Finally, the availability of forward and reverse genetic strategies in both the host and the invader will permit the development of mutants to test the importance of the genes identified as being involved in these interactions.

Abiotic Interactions

The individual plant species of today are the results of evolution in response to a multitude of biotic and abiotic environmental variations. The ways that plants have coped with the number of biotic stresses have been considered above. However, plants also encounter a myriad of abiotic stressful environmental conditions including heat, drought, salt and various metals such as aluminum, as well as variations in light. Thus plants under stress integrate a diverse range of environmental and metabolic signals through a network of signal transduction pathways that function to regulate changes in gene expression.

If the plant is growing under optimal conditions, cellular homeostasis is achieved through the coordination of a wide variety of biochemical pathways. If the plant is growing under suboptimal conditions, generally termed stress, then the integration of the various pathways that normally achieve cellular homeostasis can be disrupted, because these pathways may themselves be differentially affected by the stressful conditions. This disruption is frequently accompanied by the formation of reactive oxygen species (ROS). Because many stress responses are mediated through a response to ROS, plants make use of common pathways that allow them to acclimate to a range of different stresses, irrespective of the initiating event. Therefore, to provide adequate protection against such a hazardous environment a common signaling system has evolved, known as cross-tolerance (Bowler and Fluhr, 2000).

Any investigation of these stress responses in plants must also be associated with an understanding of the physiology of the organism. In addition, it is insufficient to characterize the responses simply at the transcriptionally level because it is clear that many of the important physiological modifications occur through the activation of protein kinases, and therefore the proteome also must be characterized.

Many of the stress responses in plants have been investigated individually. However, under normal conditions the plants may experience a combination of stresses that may result in changes somewhat different from those observed in response to any single stress. Therefore, the understanding of plant responses to more complex combinations of perturbations is important especially with reference to the applicability of the conclusions to crop improvement. Such combinations of stresses frequently occur in parts of the

world where sophisticated genomics technology is only just beginning to make an impact, and where food security issues are of paramount importance.

The understanding of abiotic stress tolerance, and the breeding for stress tolerance, have proved difficult because of the large number of genes apparently involved. The lack of traditionally bred plant lines that are multiply stress tolerant and high yielding, itself another complex trait, is a consequence of these difficulties. The application of genomics techniques to the problem has resulted in a large amount of data, especially concerning transcriptional regulation (Bohnert et al., 2001).

Drought Stress

Around the world water availability is probably the dominant environmental factor that limits plant productivity, so the search for drought tolerance is of vital importance. Drought-tolerant plants are, by definition, able to maintain the water content of their tissues or able to survive a reduction in tissue water content or recover when the drought stress is relieved. Plant survival under drought conditions can be achieved by two very different mechanisms. The strategy can be based on drought avoidance, whereby the plant architecture avoids experiencing the stress by, for example, developing a deep root system. Alternatively, the tolerance can be achieved through mechanisms by which the plant mitigates the effects of the drought stress and is able to recover when water again becomes available. Drought stress usually results in the suppression of respiration and photosynthesis. Stomatal closure and osmotic adjustment are among the responses by which the plant can limit water loss.

The identification of the genes with primary roles in any of the stress tolerances requires a series of correlations to be established and then direct verification through forward or reverse genetic manipulations to confirm the function of each of the identified genes. Some of the useful correlations include:

- That between the presence of a gene and a specific phenotype
- That which establishes a role for the gene in the evolution of a particular phenotype
- That which establishes the importance of a particular mechanism, for example, protection during drying and repair on rehydration

The availability of microarrays allows the global expression patterns in response to many stresses to be determined. At the next level, the use of proteomics allows the pattern of protein variation to be characterized. A combination of the two approaches is necessary to generate a complete picture of the response of plants to this particular stress. It must also be recognized

that the degree of variation existing among the various levels of stress tolerance, and tolerance mechanisms, requires the experiments to be performed on a wide range of biological material. Thus the proteomic analysis of rice leaves during drought stress and recovery (Slaekedeh et al., 2002) detected more than a 1000 protein spots. However, of these proteins, 42 showed significant changes in abundance under stress, but 27 of them exhibited a different response pattern in the two cultivars used. Sixteen of these forty-two proteins were identified through MS (15) or by cloning of the respective cDNA (1). Eleven of the proteins were from major cellular pathways known to be responsive to drought stress including protein synthesis, photosynthesis, carbon metabolism, and oxidative stress tolerance. Novel observations included the upregulation of an S-like RNase homolog that was lacking the two active site histidines necessary for RNase activity.

The 16 drought-responsive rice proteins are not closely related to the 16 drought-responsive maize proteins previously identified (Riccardi et al., 1998). Whether this is due to the fact that many drought-responsive proteins are still unidentified, or that the tissues sampled were somewhat different, awaits clarification. All of these characterizations are in the early stages, and much more data are likely to be reported in the near future that may, but not necessarily, lead to a clearer understanding of the various mechanisms involved in drought tolerance and the roles of the genes responsible for those mechanisms.

Because plants do not necessarily experience these stresses in isolation from one another, it is important to characterize the responses to multiple simultaneous stresses. When this was done in tobacco, the combined effect of drought and heat shock resulted in a pattern of response, at the level of gene expression, that was somewhat different from that observed when either of these stresses was applied singly (Rizhsky et al., 2002). Physiological measurements on tobacco plants simultaneously subjected to the combination of these two stresses resulted in closure of stomata, suppression of photosynthesis, an increase in respiration, and increased leaf temperature. The genes that were expressed, as shown by transcription levels, under the combination of stresses could be split into three categories:

- Genes that had been induced during either heat shock or drought stress alone and were now suppressed under the combination of stresses
- Genes that followed the expected expression pattern based on their activity under either of the stresses singly
- Genes that were specifically induced only when the combination of stresses was applied

These results with a combination of stresses that may be more typical of the conditions that plants encounter in the field were different from those

when either of the stresses was applied individually. They were also different from previously observed responses of plants to other single abiotic stresses. Therefore, the integration of plant responses to suboptimal growth conditions resulting from a combination of perturbations may not be simply predicted from their responses to each of those stresses applied individually.

One of the goals of understanding plant responses to abiotic stresses is to develop strategies for modifying plants in order to improve productivity under such conditions. However, such strategies must be based on the appropriate body of information. In other words, the information must be obtained from plants grown in an environment that is equivalent in all, or most, respects to the conditions that the modified plant will encounter. In addition, the wide range of multiple stresses with which plants have to cope are unlikely to be alleviated by the addition of a single gene or signal transduction pathway. If such a simple fix was available, it is likely that it would have emerged through the process of natural selection in some plant species.

References

Bohnert, H. J., P. Ayoubi, C. Borchert, R. A. Bressan, R. L. Burnap, J. C. Cushman, M. A. Cushman, M. Deyholos, R. Fischer, D. W. Galbraith, P. M. Hasegawa, M. Jenks, S. Kawasaki, H. Koiwa, S. Kore-eda, B.-H. Lee, C. B. Michalowski, E. Misawa, M. Nomura, N. Ozturk, B. Postier, R. Prade, C.-P. Song, Y. Tanaka, H. Wang, and J.-K. Zhu (2001) A genomics approach towards salt stress tolerance. *Plant Physiol. Biochem.* **39**, 295–311.

Bowler, C., and R. Fluhr (2000) The role of calcium and activated oxygens as signals for controlling cross-tolerance. *Trends Plant Sci.* **5**, 241–246.

Flor, H. H. (1971) Current status of the gene-for-gene concept. *Annu. Rev. Phytopathol.* **9**, 275–296.

Glazebrook, J. (2001) Genes controlling the expression of defense responses in *Arabidopsis*—2001 status. *Curr. Opin. Plant. Biol.* **4**, 301–308.

Hammond-Kosack, K. E., and J. D. G. Jones (1997) Plant disease resistance genes. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **48**, 575–607.

Jones, J. D.G. (2001) Putting knowledge of plant disease resistance genes to work. *Curr. Opin. Plant Biol.* **4**, 281–287.

Lahaye, T. (2002) The *Arabidopsis* RRS1-R disease resistance gene—uncovering the plant's nucleus as the new battlefield of plant defense. *Trends Plant Sci.* **7**, 425–427.

Lugtenberg, B. J. J., T. F. C. Chin-A-Woeng, and G. V. Bloemberg (2002) Microbe-plant interactions: Principles and mechanisms. *Antonie Van Leeuwenhoek Int. J. Gen. Mol. Microbiol.* **81**, 373–383.

Marie, C., W. J. Broughton, and W. J. Deakin (2001) *Rhizobium* type III secretion systems: legume charmers or alarmers? *Curr. Opin. Plant Biol.* **4**, 336–342.

Meyers, B. C., A. W. Dickerman, R. W. Michelmore, S. Sivaramakrishnan, B. W. Sobral, and N. D. Young (1999) Plant disease resistance genes encode members of an

ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J.* **20**, 317–332.

Nimchuk, Z., L. Rohmer, J. H. Chang, and J. L. Dangl (2001) Knowing the Dancer From the Dance: R-gene products and their interactions with other proteins from host and pathogen. *Curr. Opin. Plant Biol.* **4**, 288–294.

Peart, J. R., G. Cook, B. J. Feys, J. E. Parker, and D. C. Baulcombe (2002) An EDS1 orthologue is required for N-mediated resistance against tobacco mosaic virus., *Plant J.* **29**, 569–579.

Riccardi, F., P. Gazeau, D. de Vienne. and M. Zivy (1998) Protein changes in response to progressive water deficit in maize. Quantitative variation and polypeptide identification. *Plant Physiol.* **117**, 1253–1263.

Rizhsky, L., H. Liang, and R. Mittler (2002) The combined effect of drought stress and heat shock on gene expression in tobacco. *Plant Physiol.* **130**, 1143–1151.

Rustérucci C., D. H. Aviv, B. F. Holt III, J. L. Dangl, and J. E. Parker (2001) The disease resistance signaling components *EDS1* and *PAD4* are essential regulators of the cell death pathway controlled by *LSD1* in *Arabidopsis. Plant Cell* **13**, 2211–2224.

Salekdeh, G. H., J. Siopongco, L. J. Wade, B. Ghareyazie, and J. Bennett (2002) Proteomic analysis of rice leaves during drought stress and recovery, *Proteomics* **2**, 1131–1145.

Staskawitz, B., and M. Parniske (2001) Biotic interactions. Genomic approaches to interactions of plants with pathogens and symbionts. *Curr. Opin. Plant Biol.* **4**, 279–280.

Stougaard, J. (2001) Genetics and genomics of root symbiosis. *Curr. Opin. Plant Biol.* **4**, 328–335

Tornero, P., P. Merritt, A. Sadanandom, K. Shirasu, R. W. Innes, and J. L. Dangl (2002) RAR1 and NDR1 contribute quantitatively to disease resistance in *Arabidopsis*, and their relative contributions are dependent on the R gene assayed. *Plant Cell* **14**, 1005–1015.

Varet, A., J. Parker, P. Tornero, N. Nass, T. Nurnberger, J. L. Dangl, D. Scheel, and J. Lee (2002) NHL25 and NHL3, two NDR1/HIN1-1ike genes in *Arabidopsis thaliana* with potential role(s) in plant defense. *Mol. Plant-Microbe Interactions* **15**, 608–616.

CHAPTER

8

Identification and Manipulation of Complex Traits

Overview

The aim of most of the previous chapters has been to provide knowledge, in general terms, relating to plant genome structure and function and to provide a basic understanding of how transcription and translation products are characterized.

In this chapter problems associated with the identification, isolation, and manipulation of the genes underlying complex traits are considered. Single genes that have a very large effect on the phenotype are generally relatively easily manipulated, even if they are not necessarily as easily isolated. In contrast, some traits are genetically very complex and the traits themselves are difficult to evaluate. Other traits are not difficult to evaluate, but numerous genes can be involved in their control. In the latter two cases, it may not even be possible to uncover all those genes by using a single segregating population because the full range of alleles underlying the variation in the trait may not be present in the population.

Many of the genes in higher plants may not have an obvious phenotypic effect. This is apparent from the data on insertional mutagenesis in *Arabidopsis*. Many of the individual plants that have an insertion in a gene do not show dramatic phenotypic alterations. Part of the explanation may lie in the presence of multigene families providing redundant functions, in which case all the copies of that gene family must be silenced before any phenotypic effect will be seen (see Chapter 6). Alternatively, the effect of

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

many genes may be subtle and require dissection in an appropriate way before their roles can be adequately described. For example, the genes underlying stress tolerance, such as those for cold and drought stress, can be genetically mapped. Many of the loci that contribute to these responses have been mapped and manipulated in crosses. These loci, called quantitative trait loci (QTLs), are difficult to evaluate, and the genes that condition their effects are difficult to isolate. The most popular method for isolating these genes is by a genome walking strategy after their initial mapping to a specific region of a chromosome. However, this strategy depends on the availability of many genomic resources to be successful. Despite the wealth of positional information for QTLs in many plant species, very few have actually been isolated. Therefore, the nature of these QTLs remains largely unknown.

It is possible to detect and locate the loci affecting quantitative traits by the joint analysis of the segregation of marker genotypes and the phenotypic values of individuals in appropriate segregating populations. However, these QTLs are difficult to identify both because of the lack of discrete phenotypic segregation and because the phenotypic effects of each gene associated with a complex trait are usually relatively small. In practice, QTL analysis involves the selection and hybridization of parental lines that differ in one or more different quantitative traits (for example, tomato fruit shape, size, and solid content) (Grandillo et al., 1999) followed by analyses of the segregating progeny resulting in linkages between the quantitative trait loci and known DNA markers.

To characterize and manipulate such complex characters a number of resources are necessary. These include:

- High-quality molecular maps
- Appropriate crosses and the subsequent generations of segregating populations to allow mapping of the traits of interest
- Resources to determine genotype-by-environment effects of the various loci, namely, the variation in phenotype of a single genotype grown under various environmental conditions
- Transformation systems to reintroduce the genes, to test whether or not they have the function expected
- Tissue- and/or organ-specific promoters to get the appropriate expression of the genes in cases where inappropriate expression might interfere and/or enhance effects. An example is that of the expression of trehalose in rice, where different expression in various species gives conflicting results as to the stress protection provided by this molecule.

The generation of high-quality molecular maps is also a necessary step in marker-assisted selection (MAS). MAS is based on the information retrieved through the application of molecular markers to segregating populations that are simultaneously measured for a variety of phenotypic char-

acters. The markers can be used to enhance plant breeding efforts and to speed up the creation of cultivars. Of particular importance is the unmasking and incorporation of interesting wild alleles into elite germplasm. However, to be effective, MAS requires the availability of many markers so that the interval between the marker and the gene of interest is small. When this distance is sufficiently small, subsequent testing for the actual presence of the gene itself may not be necessary. Most of the currently useful markers are ones that can be converted into some kind of sequence-tagged sites so that PCR-based methods for genotyping are available.

MOLECULAR MAPS

As mentioned in previous chapters there are a number of different types of molecular maps. The genetic map includes the loci ordered with respect to the frequency with which they recombine. A physical map is a linear order of the genomic sequence in some form whether it is as a set of ordered BACs or as a complete genome sequence. The most useful type of molecular map will be the result of integrating the genetic and the physical maps. When the genetic markers are used to anchor the physical contigs any locus that is genetically mapped can then be placed in a specific physical region of the genome. The ultimate example in plants of an integrated genetic and physical map is *Arabidopsis*, for which the complete genome is available. However, other examples that have large, but as yet incomplete, genomic resources are maize (http://www.maizemap.org/iMapDB/iMap.html) and soybean (http://hbz7.tamu.edu/homelinks/phymap/soybean/soytool_4.htm).

DNA-based markers have revolutionized the whole process of generating genetic maps because, for the first time, a large number of loci can be followed in a single segregating population. The range of genetic markers that are available includes restriction fragment length polymorphisms (RFLP), random amplified polymorphic DNAs (RAPD) (Williams et al., 1990), amplified fragment length polymorphisms (AFLP) (Vos et al., 1995), single-nucleotide polymorphisms (SNPs), and simple sequence repeats (SSRs) or microsatellites (Senior and Heun, 1993). The usefulness or necessity of each of the marker systems depends largely on the genetic resources available for the species under consideration.

MARKER SYSTEMS

RFLPs

These were the first generation of markers and are generally detected by the hybridization of a probe to restriction-digested genomic DNA. Many of the

probes are derived from single-copy sequences, with EST sequences being one rich source of potential probes. One current application of RFLPs is the generation of a high-density map in wheat. In this example, a series of chromosome deletion lines are subjected to hybridization with labeled cDNAs as probes (http://wheat.pw.usda.gov/NSF/). Despite there being multiple copies of the genome present in hexaploid wheat, the different members of the homologous groups can be identified by the disappearance of a band in the deletion line even if there were no polymorphisms available for that particular gene (Figure 8.1). Thus deletion lines are exceptionally useful for mapping regions that are polymorphism poor. Unfortunately, this type of deletion resource is not available for many plant species, although a set of maize-oat addition lines that can be used for mapping maize genes has been constructed (Okagaki et al., 2001).

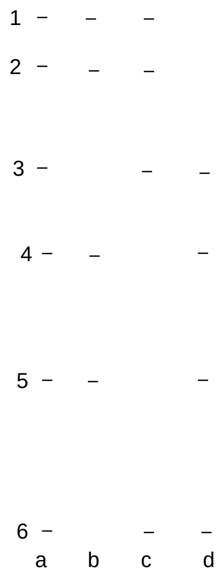


FIGURE 8.1. Using deletion lines to map chromosomal locations of ESTs. Lanes a–d represent restriction-digested DNA from a hexaploid plant (containing 6 chromosomes designated 1A, 1A, 2B, 2B, 3C, 3C) and where the DNA was extracted from: a) Plants that have all the chromosomes; b) plants that are 2B, 2B, 3C, 3C; c) plants that are 1A, 1A, 3C, 3C; d) plants that are 1A, 1A, 2B, 2B. The Southern blot was hybridized with an EST sequence that showed no polymorphisms in any of the accessions and so could not be mapped. However, from the use of deletion stocks it is possible to place the specific EST: bands 1 and 2 are on chromosome 3, bands 4 and 5 are on chromosome 2, and bands 3 and 6 are on chromosome 1. Deletion stocks that have less than a whole pair of chromosomes missing can be used to place the ESTs more accurately on the chromosomes.

AFLPs

These markers are RFLPs detected by PCR amplification. The polymorphic fragments are observed against the background of all of the possible sized restriction fragments that can be amplified. Adaptors are added to the ends of restriction fragments, and these adaptors are then used as primers in a PCR reaction. Every possible band should be amplified, and the complex mixture of bands is separated on gels or through automated sequencers. The polymorphisms can be cloned and sequenced to generate sequence-tagged sites (STSs). Potential epigenetic effects resulting from hyper- or hypomethylated regions of the genome can be investigated by using methylation-sensitive and -insensitive restriction enzyme isoschizomers. However, as mentioned in Chapter 1 species with large amounts of DNA (>20 pg per 1C) can be problematic when studying genetic diversity with AFLP techniques.

RAPDs

Statistically a sequence of 10 bp should appear once every 10^6 nucleotides. PCR amplification using genomic DNA as the target and a series of single random 10-base primers has been very successful in generating large numbers of polymorphisms (Williams et al., 1990). The methodology can be used when little other genomic information is known. Unfortunately, the technique appears to suffer from irreproducibility between laboratories and sources of thermostable enzyme, although, within a laboratory, reproducible results can be achieved (Jones et al., 1997).

Microsatellites and SSRs

Microsatellites or SSRs are genetic markers that are derived from short (usually <6 bp) tandemly repeated sequences such as $(GA)_n$, $(AAT)_n$, $(GT)_n$. The terms are often used interchangeably, although microsatellites are generally longer than the 2- to 3-bp unit of the SSRs. This type of sequence is widely dispersed through most animal and plant genomes and polymorphisms are due to the variability in the number of repeats at a given site. SSRs can be isolated from genomic libraries or enriched genomic libraries (Panaud et al., 1995) or generated from an analysis of cDNA sequences. They can be converted into STSs with primers designed in unique regions surrounding the repeat and have become an important source of genetic markers for many eukaryotic genomes (Panaud et al., 1995), especially when the primers are designed in a conserved region of a transcribed sequence, making them applicable over a wide range of taxa.

Single-Nucleotide Polymorphisms

Single-nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence differs between two individual DNA samples. For example, a SNP might change the DNA sequence AGGATTCA to AGGATTTA. SNPs can occur in both coding (gene) and noncoding regions of the genome. Many SNPs have no effect on cell function because they may not change protein structure (in fact, any SNP that occurs at the third position in the amino acid codon will have no effect if it does not change the amino acid sequence of the resulting protein). Their high frequency (perhaps as high as 2–3% in plant DNAs) means that they can be particularly useful in linkage mapping (Kristensen et al., 2001; Lai, 2001). They must be derived from sequence information, and that information must be obtained from two or more individuals. Informatics tools can be used to compare the sequences and identify variations, but the raw data in the form of trace files may be important in deciding which polymorphisms may be real. In the building of unigene sets, these differences are eliminated in the formation of the consensus sequence and need to be retrieved. Because there is no *a priori* way of differentiating between a true SNP and sequencing errors, each potential SNP must be validated. Even at a frequency of 1% these polymorphisms would generate an exceptionally large number of haplotypes if every polymorphism could be inherited independently. However, relatively few haplotypes are observed, indicating that perhaps the rate of SNP production is similar to the rate at which recombination occurs across the regions of the genome making up the haplotype blocks. Therefore, SNPs are most likely to be useful for defining haplotypes, rather than for their information individually, and so the use of SNPs is likely to involve linkage disequilibrium studies using the haplotype rather than the use of specific SNPs as individual molecular markers.

QTL Identification

Traits such as flowering time, seed development, and defense response have been extensively studied in the context of germplasm diversity, environmental adaptation, and artificial selection, and QTLs associated with such key agronomic characteristics in a wide range of cultivated and wild germplasm have been identified. The identification of QTLs from wild relatives of crop plants may be of particular importance because they may represent new opportunities to improve plant performance, with their effects on performance unpredictable from the phenotypes of the parents.

The hypothetical data shown in Figure 8.2 illustrate how QTLs for flowering could be mapped and isolated.

The initial identification of QTLs can be achieved by using either segregating populations or near isogenic lines (NILs). In both cases the initial

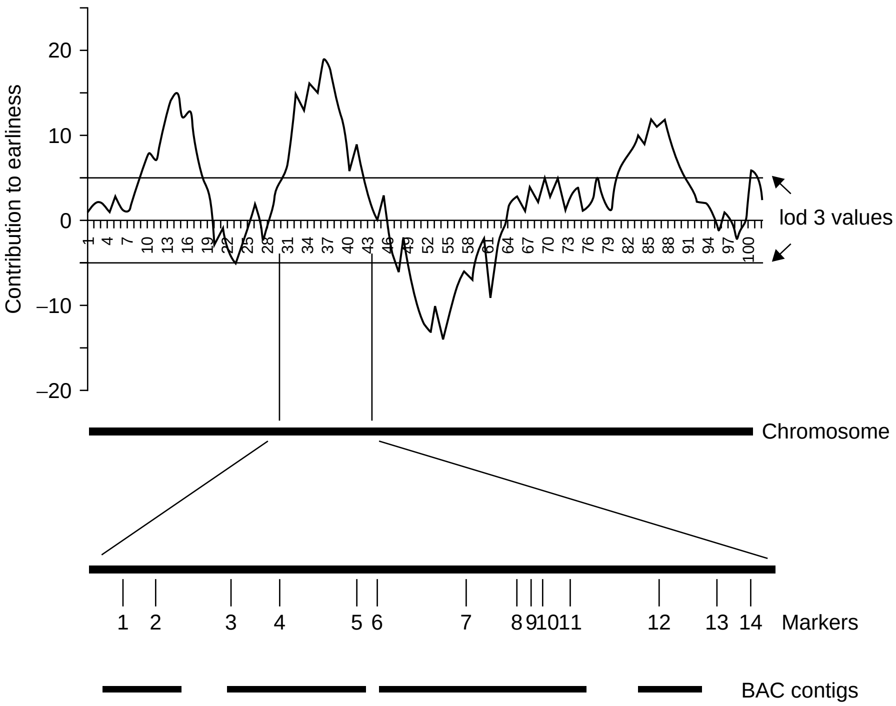


FIGURE 8.2. General scheme for the identification and positional cloning of QTLs. For a hypothetical species, molecular markers evenly spaced at 2 cM are available. A cross between two parents (each homozygous at all loci) differing in flowering time is made and the F1 backcrossed to the later-flowering parent (BC1). The BC1 generation is typed for flowering time, and the presence of a segment from the early-flowering parent for each marker is scored. The horizontal lines represent the cutoff for significance that depends on the stringency and the size of the population used. The proportion of the early lines that contain the chromosomal marker for the early parent are shown on the y-axis, and the chromosome position is shown on the x-axis. If a chromosomal region is neutral for the trait under consideration, then the two parental alleles in that region should be equally represented.

experiments involve typing the individuals for both their phenotype as well as their genetic constitution with respect to the two parents that were used to construct the population. The data in Figure 8.2 are hypothetical for mapping of QTLs that affect the flowering time. Individuals from a backcross population (BC1) between an early-flowering line (P1) and a late-flowering line (P2) [(P1 × P2) × P2] were scored for flowering time. One hundred individuals that were early flowering were genotyped to determine the frequency with which each of the parental genotypes was present in these early-flowering individuals. The two horizontal lines indicate where a

likelihood of detection (Lod) = 3 falls, indicating a significant difference from the expectation of equal representation from both parents if no genes affecting flowering time reside in that region. The analysis of the data indicates four major possible QTLs located between map positions 12–18, 31–45, 52–59, and 80–90 cM. Three of these are positive QTLs (12–18, 31–45, and 80–90 cM) where the presence of that region from P1 reduces the time to flowering. The fourth (52–59 cM) delays flowering time (even though it is derived from the early-flowering parent). However, there are three other possible QTLs between positions 20–25 (another negative one), 67–75, and 99–102. Each of these regions must have additional mapping information to reduce the distance that needs to be covered by the chromosome walking step. This additional information can be generated either from typing additional individuals from the original population, or from subsequent generations, for example, the F3 or further backcross generations and focusing on those individuals that have already been shown to be early flowering. The fine mapping of the region containing the QTLs will probably require the development of additional genetic markers before the region is be saturated. The source of these molecular markers is frequently SSRs that have been identified from sequences available for the region in question. These sequences can be either from the species under investigation or from a related species, by using previously determined syntenic relationships. As mentioned above, it is important to narrow down the size of the genetic interval before any chromosome walking takes place. Once again, the disparity in genetic resources across various crop plants would become evident. It would be relatively easy to fine map the QTL region and identify the genes in *Arabidopsis* and rice because their complete genomic sequences are available. However, this will not be the case for other crop species where the knowledge of the entire genome is still incomplete, and so it would be much more difficult to increase the number of genetic markers across the region. Once again, fine mapping of the region in question would be facilitated by utilizing the information gained from the genomic resources from other species. An example of this utilization would be the incorporation of information derived from syntenic comparisons such as those from the use of overgo oligos. Other examples of these resources have been described in earlier chapters.

In all the studies used to characterize QTLs the biological material will be of prime importance. The main focus of the information contained in this text is on the genomics and molecular techniques, but, for these to give useful information, they must be applied to an appropriate set of biological material. In particular, the development of biological material could well be the rate-limiting step in the identification and isolation of the genes underlying QTLs. Thus one could probably sequence the genome of a pine tree more rapidly, and certainly at greater expense, than one could generate a useful NIL mapping population for any particular QTL.

There are three main requirements for the fine mapping of QTLs:

1. The availability of NILs that contain the segment under investigation. Although the F2 population can be used, the NILs have the advantage of only containing the chromosomal segment that varies and so avoids the problem of background noise from other regions of the genome.
2. A large segregating population of the NILs to narrow the region, by recombination, on which the QTL resides.
3. A target region that is saturated with markers. These high-resolution maps are currently based on SSRs, but SNPs are becoming more widely used as additional high-quality genomic sequences become available. However, these additional markers still have to be generated, not a trivial task in itself.

One of the many problems that can be encountered when trying to fine map QTLs is the possibility that the phenotype being measured does not accurately reflect the genotype of the individual. Possible reasons for this lack of association are that the phenotypes of individual plants are affected by the environment or genetic background to such an extent that the mapping is confounded. These considerations are much more important when the QTLs give relatively small effects and/or the traits are difficult to measure under controlled growth conditions.

Once the fine map of the loci has been completed, the physical isolation of those regions must be undertaken. This can be relatively simple if the region has already been covered in one of the physical mapping projects and the appropriate clones are readily available. In general, the identification of the regions will be based on the markers and a search of the appropriate databases to determine what resources are already available. Possible resources would be a BAC library or a series of contigs that include the markers in question. However, if these physical resources are not available then the BAC library will have to be generated and subsequently screened to isolate the target regions. These regions then must be sequenced and the sequence interrogated to identify any putative genes that may be present. The putative genes would be identified from predicted open reading frames with gene finding programs such as those described in Chapter 9. The predicted open reading frames would be used in database searches to determine possible relationships with known genes. Any genes that are known to function in pathways that may be associated with the QTLs' effect are likely to be the first targets of investigation.

An alternative to increasing the population sizes to narrow the interval in which the gene conditioning the QTL lies, is to take advantage of historical meiotic events by association or linkage disequilibrium studies (Jannink et al., 2001; Jansen and Nap, 2001; Nordborg and Tavare, 2002). Therefore,

in unrelated individuals that differ for the trait under consideration, the markers associated with the QTL can be assayed. Those in which there is no association with the trait can be excluded, and in this way the region containing the QTL can be reduced.

Confirmation of the Genes Conditioning the QTLs

Because it is impossible to narrow down the region containing the QTLs to a single gene by screening very large numbers of individuals in a segregating population, it is important to confirm the function of the candidate gene(s). The most direct way of doing this would be by transformation and complementation analysis. The candidate gene can be reintroduced into the appropriate genotype by using a number of different constructs:

- As a cDNA under the control of a strong promoter
- As a short stretch of genomic DNA to include the endogenous promoters
- As a large segment such as a cosmid or BAC clone because the trait may be controlled by a number of very closely linked genes or by important regions some distance away from the candidate gene

A plant-transformation-competent BIBAC library would be a useful resource in the confirmation of the identity of QTL regions. This library would permit a large segment containing the putative QTL to be introduced. Such a library has been constructed for the *Arabidopsis thaliana* Landsberg ecotype (Chang et al., 2003). However, the specific alleles that are responsible for the QTLs are not all located in a single variety, and therefore many libraries would have to be constructed, one from every one of the lines that contain interesting QTLs.

A Practical Example of QTL Isolation—the fw2-2 Locus in Tomato

Frary et al. (2002) have described in detail the cloning of a gene underlying a QTL in tomato. The *fw2-2* locus is involved in the control of tomato fruit size. This body of work demonstrated the level of additional biological and genetic information necessary for the positional cloning approach to be successful for QTL gene isolation (Grandillo et al., 1999; Alpert and Tanksley, 1996).

Fruit size and shape are major determining factors related to the yield, quality, and consumer acceptability of fruit crops. These two traits are inherited in a quantitative fashion and have been extensively studied in tomato (Figure 8.3). The major QTLs responsible for these traits have been identified (Grandillo et al., 1999). These QTL studies were conducted over a period

Publisher's Note:
Permission to reproduce this image online was not granted by the copyright holder. Readers are kindly requested to refer to the printed version of this article.

FIGURE 8.3. (**A**) Fruit size extremes in the genus *Lycopersicon*. On the left is a fruit from the wild tomato species *L. pimpinellifolium*, which like all other wild tomato species, bears very small fruit. On the right is a fruit from *L. esculentum* cv Giant Red, bred to produce extremely large tomatoes. (**B**) Phenotypic effect of the *fw2.2* transgene in the cultivar Mogeor. Fruit are from R1 progeny of fw107 segregating for the presence (+) or absence (–) of cos50 containing the small-fruit allele. (Reprinted with permission from Frary et al., Science **289**, 85–88. Copyright (2002) American Association for the Advancement of Science).

of two decades and included many populations that were segregating for fruit size. Through these studies at least 28 QTLs have been identified. However, not all of the QTLs have been fine mapped, and it is possible that some may contain more than one gene. Of these 28, the locus *fw2.2* had the greatest effect on fruit size, contributing up to 30% of the variation in fruit weight.

The quantitative trait locus *fw2.2* was mapped to the tomato chromosome 2. A high-resolution physical and genetic map of the region contain-

ing this locus was constructed using an F2 nearly isogenic line mapping population (3472 individuals) derived from *Lycopersicon esculentum* (domesticated tomato) × *Lycopersicon pennellii* (wild tomato) population (Alpert and Tanksley, 1996). The genetic size of this region was found to be 0.13 cM, and the physical distance was estimated to be less than 150 kb by pulsed-field gel electrophoresis of tomato DNA. A physical contig of the region composed of six yeast artificial chromosomes (YACs) was isolated. The mapping population used resulted in the placement of *fw2.2* in a region of less than 150 kb, which was completely included in two YACs (Figure 8.4). The fact that the phenotypic effect of the *fw2.2* QTL could be mapped to such a small interval is consistent with the notion that this particular QTL is likely to be due to a single gene.

One of the YACs was used to screen a cDNA library (from *L. pennellii*), and four transcripts were identified. The four cDNAs were then used to screen a cosmid library (again from *L. pennellii)* from which four positive and nonoverlapping cosmids, one corresponding to each unique transcript, were isolated. Each of these cosmids was transformed into two different tomato varieties, each of which contained a partially recessive large fruit allele of fw2.2. Because the fw2.2 *L. pennellii* allele is only partly dominant, and the primary transformants were only hemizygous for the transgene, the primary transformants were self-pollinated. The progeny fell into three classes, homozygous, hemizygous, or null for the introduced gene.

One of the cosmids (cos 50) showed differences in two independent transformants for the presence or absence of the transgene (Figure 8.3). Because the transformations were not the result of homologous recombination with the replacement of the endogenous gene with the transgene, the fact that two independent transformants show an effect is consistent with the conclusion that the gene functions similarly in various genomic locations.

The sequence analysis of cos 50 revealed two open reading frames. One of these open reading frames, the cDNA 44 which was originally used to

FIGURE 8.4. High-resolution mapping of the *fw2.2* QTLs. (**A**) The location of *fw2.2* on tomato chromosome 2 in a cross between *L. esculentum* and a NIL containing a small introgression (gray area) from *L. pennellii.* (**B**) Contig of the *fw2.2* candidate region, delimited by recombination events at XO31 and XO33. Arrows represent the four original candidate cDNAs (70, 27, 38, and 44), and heavy horizontal bars are the four cosmids (cos62, 84, 69, and 50) isolated with these cDNAs as probes. The vertical lines are positions of restriction fragment length polymorphism or cleaved amplified polymorphism (CAP) markers. (**C**) Sequence analysis of cos50, including the positions of cDNA44, ORFX, the A-T-rich repeat region, and the "rightmost" recombination event, XO33. (Reprinted from Frary et al., Science **289**, 85–88. Copyright (2002) American Association for the Advancement of Science).

Publisher's Note:
Permission to reproduce this image online was not granted by the copyright holder. Readers are kindly requested to refer to the printed version of this article.

isolate cos 50, was shown by genetic analysis not to be involved in the QTL (Figure 8.4). The open reading frame X (ORFX) was transcribed at very low levels and had homologs in other plant species and a predicted structural similarity to the human oncogene RAS protein. Therefore, ORFX was analyzed to determine whether it was the basis of *fw2-2*. The sequence analysis of ORFX and upstream regions indicated that the changes in phenotype caused by various alleles of *fw2.2* were not solely due to differences within the coding region of ORFX but could be modulated by a combination of sequence changes in both the coding and upstream regions of *ORFX*. A comparable situation exists in maize, where variation in the upstream regulatory regions of the *teosinte branched1* gene has been implicated in its domestication (Wang et al., 1999).

Thus the isolation of *fw2.2* was dependent on all the following resources:

- A QTL of fairly large effect
- A number of different mapping populations segregating for the QTL that enabled its initial detection
- A large NIL segregating population for the fine mapping of the QTL
- A YAC library that covered the QTL region
- The appropriate cDNA library for the identification of the transcribed regions in the YAC
- A cosmid library from the appropriate line for the transformation experiments
- Transformable varieties for testing the putative QTL regions
- A QTL allele used for transformation that was at least partially dominant (because homologous recombination is not available to remove the endogenous copy of the gene)
- Some luck, because the actual open reading frame was not one identified in the original cDNA library screen

These resources will not be available for large number of different crop species. However, the mapping of QTLs can still be important and useful, and their manipulation in MAS schemes greatly facilitates the development of new varieties.

Marker-Assisted Selection

Traditional plant breeding is based on the selection of plants based on visible or measurable traits, the phenotype. Marker-assisted selection (MAS), on the other hand, can be used provided the molecular markers are sufficiently close to the gene under selection. A standard backcross breeding scheme is shown in Figure 8.5, and that for a MAS is shown in Figure 8.6.

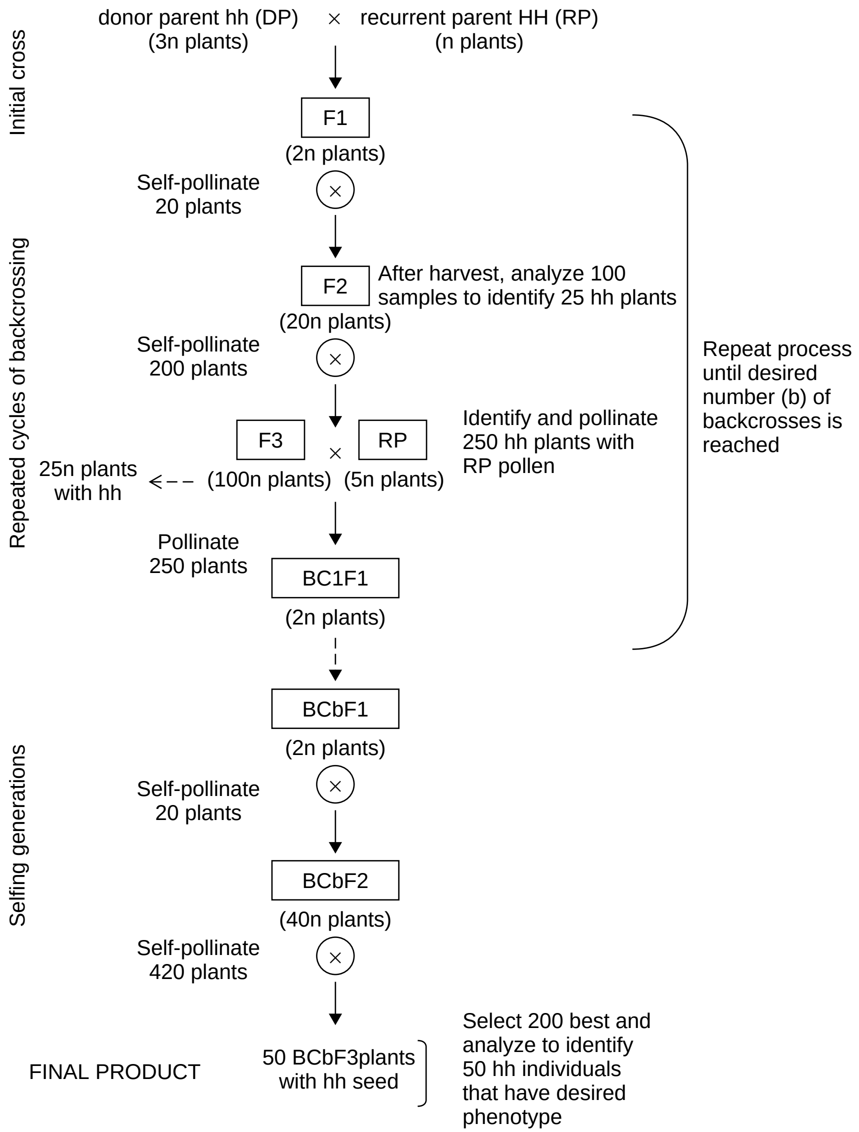


FIGURE 8.5. Conventional backcross breeding scheme. (Adapted from Dreher et al., 2000.)

Conventional breeding methods have produced impressive genetic gains, but it is clear that this historical rate of improvement cannot be sustained using these methods alone (Dreher et al., 2000; Dreher et al., 2003; Morris et al., 2003). Therefore, the tools that have become available through

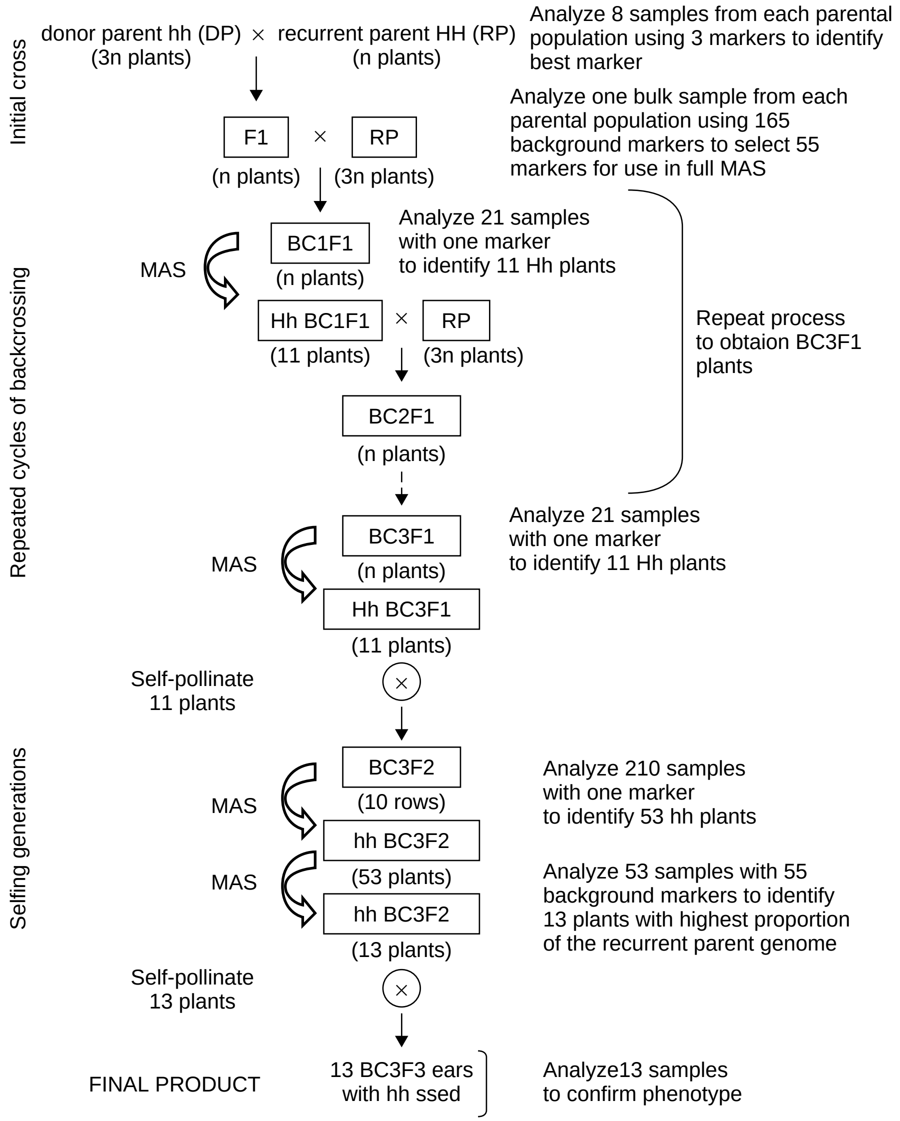


FIGURE 8.6. MAS backcross breeding scheme. (Adapted from Dreher et al., 2000).

genomics will extend the repertoires available and complement conventional breeding. MAS has the potential to increase the efficiency of a breeding scheme by:

- Reducing the time required to develop a new variety
- Lowering the size of populations needed, thereby eliminating large and costly field evaluation

Conventional Breeding Methods

Conventional breeding methods involve three basic steps:

- The production of a population of plants incorporating the desirable traits
- The evaluation and selection of the best within this population
- The intercrossing of these superior individuals

The new population is then passed through subsequent cycles of selection and improvement.

What happens for a trait that is only expressed in a seed, but does not have a visible phenotype, for example, protein quality, in a conventional breeding program? Because this is a seed trait the breeders cannot identify individuals that have an improvement until the seed has been tested. Therefore, they either have to wait until the end of the season to identify the desirable plants or carry through a much larger number of plants, many of which will subsequently turn out to be useless. Because the trait will need biochemical analysis, the superior individuals cannot be identified until some time after the seeds are harvested. If the gene that is being introduced is recessive it will not be expressed in the first generation, so the heterozygotes cannot be identified. Therefore, the individuals also must be self-fertilized, and a small number of progeny must be typed to identify the heterozygotes. These cycles must be repeated to develop a new line with the introduced gene in a genetic background as similar to the original commercial line as possible.

Implications of MAS. Even if the plant's phenotype under selection is a reliable indicator of the underlying genetic characteristic, the phenotypic evaluation can be costly, time consuming, and affected by the growth environment. The use of molecular markers has the potential to provide a solution to these problems. Provided they are sufficiently closely linked to the desirable allele, the presence of such an allele can be determined directly in seedlings by the use of these markers. This eliminates the costly and time-consuming phenotypic evaluation. The molecular markers can also be used to distinguish between the homozygous and heterozygous plants, thereby eliminating the need to self individuals to determine their genetic makeup. An added benefit of MAS arises when a large number of molecular markers that cover the plant's entire genome are used. These markers can then identify those individuals that contain the largest contribution of genetic material from the recurrent recipient line, thereby reducing substantially the number of generations required to introgress the desired allele.

A comparison of conventional breeding methods and MAS has been carried out (Dreher et al., 2000; Dreher et al., 2003; Morris et al., 2003). The

study identified a number of areas in which MAS would offer significant advantages over conventional breeding methods. These advantages are:

- A reduction in the extent of phenotypic screening
- The ability to identify the presence of multiple alleles related to a single trait even if the alleles do not individually produce a detectable influence on the expression of the traits
- The ability to select multiple traits simultaneously
- The screening of traits whose expression depends on the growth environment
- The early detection of superior lines, especially for seed-expressed traits
- The ability to manipulate recessive genes and identify the heterozygotes
- A reduction in the number of backcrossing cycles

Molecular markers are clearly a powerful tool for plant breeding. MAS offers opportunities for reducing costs, saving time, and accomplishing breeding goals unavailable through conventional methods. However, there have been different levels of adoption of the MAS strategy between commercial and publicly funded breeding efforts. Part of the explanation for the relatively slow integration of MAS is the expense of establishing biotechnology research facilities, the difficulty of identifying useful markers, and the identification of markers linked to traits of interest that are controlled by a large number of minor genes. The continued development of new technology, including DNA chip technology and SNP detection, should overcome some of these constraints. Especially important will be the ability to study the expression of thousands of genes related to a trait of interest simultaneously and to dissect out that subset which is causative rather than consequential. The identification of the genes that play dominant roles in plant growth and development will also help narrow the focus in the search for genes of importance in agronomic improvement.

References

Alpert, K. B., and S. D. Tanksley (1996) High-resolution mapping and isolation of a yeast artificial chromosome contig containing fw2.2: A major fruit weight quantitative trait locus in tomato. *Proc. Natl. Acad. Sci.* USA **93**, 15503–15507.

Chang Y. L., X. Henriquez, D. Preuss, G. P. Copenhaver, and H. B. Zhang (2003) A plant-transformation-competent BIBAC library from the *Arabidopsis thaliana* Landsberg ecotype for functional and comparative genomics. *Theoret. Appl. Genet.* **106,** (2): 269–276.

Dreher, K., M. Morris, M. Khairallah, J-M Ribaud, S. Pandey, and G. Srinivasan (2000) Paper presented at the Fourth Annual Conference of the International Consortium on Agricultural Biotechnology (ICABR), "The Economics of Agricultural Biotechnology," Ravello, Italy, 24–28 August, 2000.

Dreher, K., M. Khairallah, J. M. Ribaut, and M. Morris (2003) Money matters (I): costs of field and laboratory procedures associated with conventional and marker-assisted maize breeding at CIMMYT. *Mol. Breeding* **11**, 221–234.

Frary, A., T. C. Nesbitt, A. Frary, S. Grandillo, E. Van Der Knaap, B. Cong, J. P. Liu, J. Meller, R. Elber, K. B. Alpert, and S. D. Tanksley (2000) Fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**, 85–88.

Grandillo, S., H. M. Ku, and S. D. Tanksley (1999) Identifying the loci responsible for natural variation in fruit size and shape in tomato. *Theoret. Appl. Genet.* **99**, 978–987.

Jannink, J. L., M. C. A. M. Bink, and R. C. Jansen (2001) Using complex plant pedigrees to map valuable genes. *Trends Plant Sci.* 6, 337–342.

Jansen, R. C., and J. P. Nap (2001) Genetical genomics: the added value from segregation *Trends Genet.* **17**, 388–391.

Jones, C. J., K. J. Edwards, S. Castaglione, M. O. Winfield, F. Sala, C. vandeWiel, G. Bredemeijer, B. Vosman, M. Matthes, A. Daly, R. Brettschneider, P. Bettini, M. Buiatti, E. Maestri, A. Malcevschi, N. Marmiroli, R. Aert, G. Volckaert, J. Rueda, R. Linacero, A. Vazquez, and A. Karp (1997) Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Mol. Breeding* **3**, 381–390.

Kristensen, V. N., D. Kelefiotis, T. Kristensen, and A. L. Borresen-Dale (2001) High-throughput methods for detection of genetic variation. *Biotechniques* **30**, 318–321.

Lai, E. (2001) Application of SNP technologies in medicine: Lessons learned and future challenges. *Genome Res.* **11**, 927–929.

Morris, M., K. Dreher, J.-M. Ribaut, and M. Khairallah (2003) Money matters (II): costs of maize inbred line conversion schemes at CIMMYT using conventional and marker-assisted selection. *Mol. Breeding* **11,** 235–247.

Nordborg, M., and S. Tavare (2002) Linkage disequilibrium: what history has to tell us *Trends Genet.* **18**, 83–90.

Okagaki, R. J., R. G. Kynast, S. M. Livingston, C. D. Russell, H. W. Rines, and R. L. Phillips (2001) Mapping maize sequences to chromosomes using oat-maize chromosome addition materials. *Plant Physiol.* **125**, 1228–1235.

Panaud, O., X. L. Chen, and S. R. McCouch (1995) Frequency of microsatellite sequences in rice (*Oryza sativa* L). *Genome* **38**, 1170–1176.

Senior, M. L., and M. Heun (1993) Mapping maize microsatellites and polymerase chain-reaction confirmation of the targeted repeats using a CT primer. *Genome* **36**, 884–889.

Vos, P., R. Hogers, M. Bleeker, M. Rijans, T. Van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau (1995) AFLP—A new technique for DNA-fingerprinting. *Nucleic Acids Res.* **23**, 4407–4414.

Wang, R. L., A. Stec, J. Hey, L. Lukens, and J. Doebley (1999) The limits of selection during maize domestication. *Nature* **398**, 236–239.

Williams, J. G. K., A. R. Kubelik, J. Livak, A. J .A. Rafalski, and S. V. Tingey (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic-markers *Nucleic Acids Res.* **18**, 6531–6535.

Web Resources

http://www.maizemap.org/iMapDB/iMap.html

http://hbz7.tamu.edu/homelinks/phymap/soybean/soytool_4.htm

http://wheat.pw.usda.gov/NSF/

CHAPTER

9

BIOINFORMATICS

OVERVIEW

Bioinformatics covers the whole range of activities from the documentation of the genomics and associated data, to its storage and analysis, and its final dissemination. The importance of bioinformatics is a consequence of the generation of very large amounts of data being produced by the genomics revolution. Coping with these data has turned into one of the major activities for all genomic projects. In general, the advances in data gathering, storage, and distribution technologies have far outpaced computational advances in techniques for analyzing and understanding data.

Experimental approaches are designed to discover genes and their functions by combining genomic sequences, ESTs, proteomics, genetics, and reverse genetics. Biological functions also must be assayed to understand genotype/phenotype correlations, and these observations must be connected to all the genomics data. A frequent additional consideration is that many genomics projects involve more than a single institution and therefore require the frequent interchange of the raw data between the various members involved in the project. The volume and complexity of all of these data generally require an organized information management system to enter, maintain, and analyze this data, so that the cycle from experiment, to analysis, to prediction, to the experimental testing of any new insights is efficiently managed.

The components that are involved in the assembly and analysis of all of these data include:

- A database (now necessarily a relational database) to store and cross-query all the project-generated data at an appropriate level of detail

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

- The software to conduct the data type-specific analyses
- General-purpose multivariate statistics to test the significance of any tentative conclusions

Although these components are considered separately here, they are frequently present together, especially in the large public databases where the data and the analysis tools are available from the same site.

DATABASES

The necessary components of a database's structure are covered to some extent in Chapter 2. The large volume of data must be organized in an informative and relational manner to allow it to be easily linked with other relevant data sets. The data sets also need to be appropriate to facilitate automated data mining, which is the extraction of hidden predictive information from databases.

The data, and the tools for mining and analyzing the data, can be made available either through a general data warehouse, for example, the National Center for Biotechnology Information (NCBI), or one focused on a specific species or set of species, examples of which are The A*rabidopsis* Information Resource (TAIR), Gramene, ZmDB, MaizeDB, and the Legume Information System.

The ultimate goal of data management is to provide data access, data mining, and modeling support. Data mining, also known as knowledge discovery in databases, uses sophisticated statistical analysis and modeling techniques to uncover predictive patterns and relationships hidden in organizational databases—patterns that ordinary methods might miss. This activity usually includes a combination of machine learning, statistical analysis, modeling techniques, and database technology.

The information necessary before setting up a database includes:

- The types of data that will be collected and stored—gel images, sequence data including trace files, phenotypes, genetic maps, etc.
- The metadata associated with each form of stored data to be included:
  a. That describing the collected data—all the necessary experimental information relating to the stored images, the growth conditions, and the developmental stages of the sampled tissue for RNAs used in microarray experiments, etc.
  b. That describing the derived data—the version of analysis programs that have produced the reported alignments, the versions of the databases included in a clustering of ESTs, etc.

If the data are used in a subsequent analysis that modifies the previous conclusions, both the original and the new conclusions should be stored and

versioned, so a returning user can trace the changes from his or her previous download. In this way subsequent searches will be interpretable by outside investigators.

Therefore, any specific genomics database must be developed so that it is compatible with:

- The existing data generation and reporting systems
- The databases to be queried
- The target database for dissemination of results

The database must also be designed to enable the analysis of the data with respect to multiple dimensions that would include the integration of variations associated with development, tissue type, genotype, and growth conditions, among other possibilities, that is, cross-queries must be facilitated both within and across databases.

Clearly, the ultimate objective of the data is to provide the basis for further investigations, and therefore the data must be available to the community of scientists who may be able to make use of it. The need to access and download data may extend from a single item to a bulk download, for example, of complete EST collections, so that the database structure must be sufficiently flexible to accommodate such diverse needs. In addition to the data accessibility concerns is the problem of defining a common vocabulary applicable to the data. Thus, if various groups use different terms to describe the same data, then the comparison of those data sets becomes much more difficult, especially with respect to automated queries. Therefore, the efforts at developing gene ontologies, for example, the Gene Ontology (GO) project, have the goal of producing a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing (http://www.geneontology.org/) (Ashburner et al., 2000).

Once the data have been generated in a suitable format, then some or all of it must be stored. Two categories of data, static and dynamic, can be archived. Examples of static data, or data that do not change frequently, are sequences, publication records, and germplasm descriptions. Dynamic data, in contrast, undergo frequent changes and comprise mainly derived data such as sequence similarity and genetic or physical maps.

Three possible levels of data storage are:

- Individual laboratory or project databases
- Specialized databases
- Central public or private databases and archives

The individual databases are databases that manage data from individual groups or projects as they are generated. Here all the detailed information

for those specific data is stored. Much of the data in these databases is likely to be dynamic data, with the static data being transferred to specialized or public databases.

Specialized databases are frequently either organism-specific databases or databases focused on a few closely related species. Examples of these specialized databases are TAIR, Gramene, ZmDB, MaizeDB, and the Legume Information System. They are likely to contain both some of the static data and the dynamic data because they also function as a central resource for that group of organisms. They are designed to house the most current and accurate data and establish standards for data exchange for complex, dynamic data types appropriate to that community. They are also restricted to a subset of all the genomics and genetics data available, because they do not store the totality of the plant genomics data, even if they do contain all the data for the representative set of organisms included. However, it is important that these specialized databases are compatible among themselves and with the wider universe of databases so that automated queries can be facilitated. The current state of genomics information and analysis has resulted in multiple versions of even the specific databases. Therefore, for *Arabidopsis thaliana* there are databases that are maintained by TAIR, MIPS, and TIGR that may have the same static data but different versions of some of the dynamic data such as the composition of the current unigene set.

GenBank is an example of a public archive, especially for nucleic acid sequence data, because it is the only place to which all such data are expected to be submitted. GenBank is expertly managed in the areas of data storage, handling, Internet access, retrieval, and analysis of nucleic acid and protein data.

Assuming that other central databases are developed for storage, the question that arises is what should be stored and for how long. These same questions arise with any specialized database as well, although the curation associated with these specialized databases can probably deal with this question adequately. However, how does the long-term storage of the data housed in individual laboratory or project databases occur, as funding (and investigators) have finite lifetimes? Once again, the importance of database standards becomes clear. Provided the structure of any boutique database is appropriate for its importation into a data warehouse, if necessary, the data can be relatively easily archived. However, the question as to who curates the public archives, and who is responsible for supporting this curation and storage, is not as easily solvable.

Informatics Tools

Once the data have been generated and appropriately stored in a database, it is usually necessary to analyze that data in some form. A large number of

tools have been, and are continuing to be, developed for this purpose to cover the whole range of genomics data that are acquired. Many of these tools are available free over the Internet, and many of the sites that house the data are also designed to allow access to the various tools so that the analyses can all be performed through that particular site. In addition, many of the tools are designed to give graphical representations/visualizations for ease of display and interpretation of the results.

INFORMATICS AND ANALYSES OF PLANT GENOMES

Genomic approaches to plant biology have provided a tremendous wealth of gene and gene sequence data for a wide variety of plant species. The availability of these large bodies of data provides both opportunities and challenges. The cross-referencing of the sequence data with the expression data available from functional genomics and proteomics projects provides an opportunity to develop a complete catalog of gene function across the plant kingdom. A start has been made on this with the 2010 program at the NSF and its counterparts in Europe and Japan. It is important to integrate diverse data and data types to provide a more complete and consistent view of all this information. Therefore, the annotation of the data must be consistent and identifiable. However, the current crop of bioinformatics software does not generate a single unambiguous conclusion, because they use different criteria either for identifying the features associated with a particular genomic sequence, for example, the identity of splice sites, or for recognizing the component open reading frames in a spliced gene.

SEQUENCE ASSEMBLY

Sequence assembly is a crucial early step in many genomics projects. This is because most of the initial sequence "reads" are very much shorter than the functional or important region of the genome being characterized. As described in Chapter 3, whole genome sequencing projects use two different approaches to generate the sequence, whole genome shotgun sequencing and minimum tiling path methods. However, in each case the sequence reads must be assembled into large contiguous regions of genomic sequence. The draft genomic sequences then must be finished. Finishing is the process of turning a rough draft assembly into a highly accurate contiguous DNA sequence with a defined maximum error rate. Therefore, finishing involves closing the remaining gaps, resolving ambiguities, and validating the assembly. The aim is that the overall sequence quality conforms to the Bermuda standard of being confirmed by at least two templates, accurate to at least 1 bp in 10,000, and with no gaps. Finishing usually makes use of multiple software tools in an iterative manner to obtain contiguous sequence that meets the standards for double-strand coverage and sequence quality

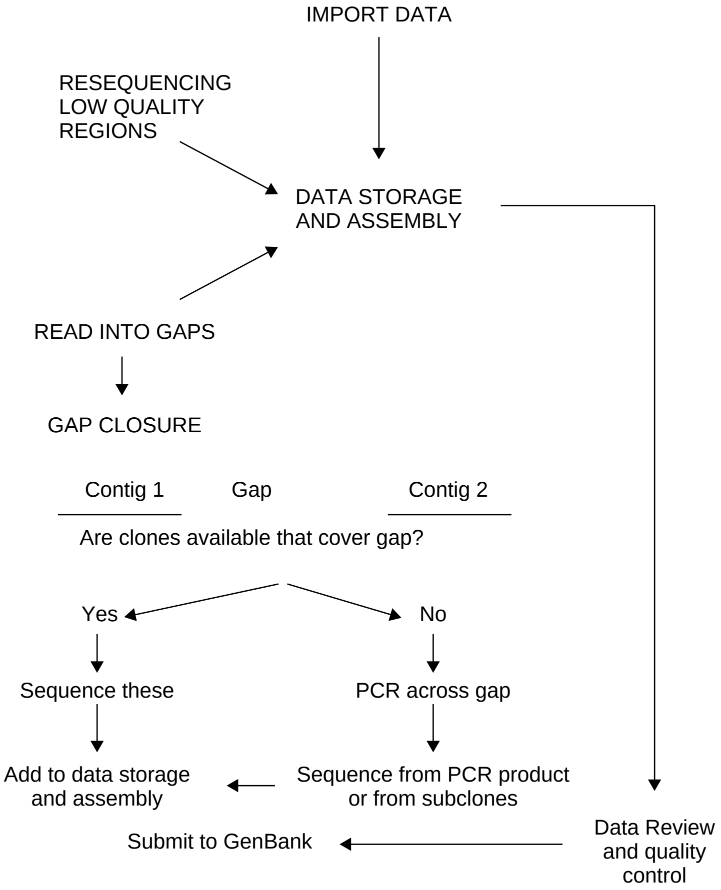


FIGURE 9.1. Genome sequence finishing. The rough draft assembly is converted into a fully assembled contiguous sequence by resequencing and gap closure (Adapted from http://www-shgc.stanford.edu/Seq/doepages/methodology.html).

(http://www.genome.gov/10001812). Once the finished assembly is complete it can be experimentally verified via PCR using selected primer pairs. These primers are often the same ones that were generated to close any gaps in the original assembly. The process of finishing a draft assembly is shown in Figure 9.1.

EST Assembly

Sequence assembly is also used to collapse the large libraries of expressed sequence tags (ESTs) to produce tentative contigs (TCs) or unigene sets (TUGs). Such assembly procedures eliminate the redundancy that exists in

the EST data because multiple ESTs may possibly represent different parts of the same gene. The clustering also can extend the length of the sequence and would ideally result in the full length of the gene being included in the final assembled product. Such collapsing of the data sets is important. Thus the 415,000 wheat ESTs in GenBank are unlikely to represent an equivalent number of separate genes given our knowledge of the gene content of other plant genomes.

The unigene set that is derived from the assembly of ESTs is not a compete set of genes for that organism because the cDNA libraries used to generate the ESTs will not reflect all the genes in a particular genome. The expression of some genes is too low or too transient to result in capture, whereas other genes will only be expressed in tissues under certain growth conditions or in particular tissues. Therefore, the unigene set is simply the least redundant set of expressed sequences that can be arrived at by using all the available data.

The extent of the coding regions included in the unigene set can be extended by the use of gene predictions from complete genomic sequences. The matching region can be aligned and gene prediction programs used to identify the possible transcription and translation start sites. This information can be used to design primers to be used in reverse transcription PCR (RT-PCR) to test whether the predicted fragments can be amplified from mRNA populations. Again, informatic analysis of the regions 5′ to the tentative expressed region may indicate where and/or when such a transcript may be expressed, thereby narrowing the range of tissues or developmental stages that must be screened to generate experimental evidence for the predicted expression. The 5′ regions are the least represented transcribed sequences, as cDNA, SAGE, or MPSS libraries are all biased toward the 3′ ends of the mRNAs because of the oligo-dT priming of the first-strand synthesis.

The Right Tool for the Right Job

Many of the database sites contain a range of tools for sequence characterization and identification. The NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) site makes available a range of tools with a regular update of all the documentation describing the programs, their uses, and the interpretation of the resulting data. Many of the specific databases also include BLAST as an integral option, although without the full range of functions or the tutorials.

BLAST

BLAST® (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regard-

less of whether the query relates to protein or nucleotide sequence. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits (http://www.ncbi.nlm.nih.gov/BLAST/blast_overview.html).

The BLAST algorithm does not only identify matches that serve to make a correct functional assignment. BLAST has a large set of variable parameters that can be altered to produce a range of different matches. The database that is the target of the BLAST search changes constantly as the available set of sequenced genes increases. This will also change the output, sometimes dramatically. Thus even a program as basic as BLAST will not generate the same results reproducibly unless users carefully coordinate and synchronize all the program's inputs. Although the NCBI site has a BLAST Program Selection Guide (reproduced in Tables 9.1 and 9.2) and a series of tutorials to aid in the use and interpretation of the data, the data still generally must be manually inspected to determine which particular matches are meaningful.

SEQUENCE MATCHING

Sequence similarity is a very general tool that forms the basis of many different biological sequence analyses. The basic tool for generating sequence matches is BLAST. It is limited by the traditional alignment presentation style of the results. An alternative program such as Miropeats (http://www.genome.ou.edu/miropeats.html) discovers regions of sequence similarity within or among any set of DNA sequences and then graphically represents the regions that are similar. An example is given in Figure 9.2, which shows the distribution of repeated segments over a short stretch of flax sequence, with a pair of inverted repeats, a tandem oligo repeat, and a palindrome. In Figure 9.3, the same region of the *Arabidopsis* chromosome 2 (BAC F16P2) that is shown in Figure 1.4 is shown as a Miropeat pattern rather than with the repeats indicated with directed arrows as in Figure 1.4. The enhancement offered by Miropeats when making conventional DNA sequence comparisons is the summary of extensive large-scale sequence similarities on a single page of graphics. Miropeats can handle the comparison of the repeat structures of entire chromosomes, visualizing overlapping sequence fragments in a contig assembly project or comparing the products of different contig assembly programs.

ANNOTATIONS OF GENOME SEQUENCE

Once the genome sequence has been assembled, the important features such as genes, transposons, and repeats, must be placed on the sequence. This

TABLE 9.1. WHICH BLAST TOOL IS APPROPRIATE FOR SPECIFIC SEARCHES?

If your sequence is NUCLEOTIDE

Length	Database	Purpose	BLAST Program
20 bp or longer	Nucleotide	Identify the query sequence	MEGABLAST (accept batch queries) Standard BLAST (blastn)
		Find sequences similar to query sequence	Standard BLAST (blastn)
		Find proteins similar to translated query in a translated database	Translated BLAST (tblastx)
	Protein	Find proteins similar to translated query in a protein database	Translated BLAST (blastx)
7–20 bp	Nucleotide	Find primer binding sites or map short contiguous motifs	Search for short, nearly exact matches

If your sequence is PROTEIN

Length	Database	Purpose	BLAST program
15 residues or longer	Protein	Identify the query sequence or find protein sequences similar to query	Standard Protein BLAST (blastp)
		Find members of a protein family or build a custom position-specific score matrix	PSI-BLAST
		Find proteins similar to the query around a given pattern	PHI-BLAST
	Conserved Domains	Find conserved domains in the query	CD-search (RPS-BLAST)
	Conserved Domains	Find conserved domains in the query and identify other proteins with similar domain architectures	Conserved Domain Architecture Retrieval Tool (CDART)
	Nucleotide	Find similar proteins in a translated nucleotide database	Translated BLAST (tblastn)
5–15 residues	Protein	Search for peptide motifs	Search for short, nearly exact matches

From http://www.ncbi.nlm.nih.gov/BLAST/producttable.html

TABLE 9.2. Specialized Databases Available from NCBI

Specialized Database Searches

Query	Database	Purpose	BLAST Program
Nucleotide or Protein	None	Compare the query and second sequence directly	BLAST 2 Sequences
	The NCBI Draft Human Genome	Map the query sequence. Determine the genomic structure. Identify novel genes.	Human Genome BLAST
	Mouse Genome	Map the query sequence. Determine the genomicstructure. Identify novel genes.	Mouse Genome BLAST
	Rat	Map the query sequence. Determine the genomic structure. Identify novel genes.	Rat Genome BLAST page
	Fugu (pufferfish)	Map the query sequence.Determine the genomic structure. Identify novel genes.	*Fugu rubripes* Genome BLAST page
	Zebrafish	Map the query sequence. Determine the genomic structure. Identify novel genes.	Zebrafish Genome BLAST page
	Arabidopsis thaliana	Map the query sequence. Determine the genomic structure. Identify novel genes.	*Arabidopsis thaliana* BLAST page
	Oryza sativa(rice)	Map the query sequence. Determine the genomic structure. Identify novel genes.	*Oryza sativa* BLAST page
	Anopheles gambiae (mosquito)	Map the query sequence. Determine the genomic structure. Identify novel genes.	*Anopheles gambiae* BLAST page
	Other eukaryotes Plasmodium, Amoeba, fungi, etc.	Map the query sequence. Determine the genomic structure. Identify novel genes.	Other Eukaryotes BLAST page
	Microbial Genomes	Map the query sequence or identify novel genes	Microbial Genome BLAST
Nucleotide	UniVec	Screen for vector contamination	VecScreen
	Trace Archives	Find matches to unassembled, raw sequence data	Trace MEGABLAST Search

From http://www.ncbi.nlm.nih.gov/BLAST/producttable.html

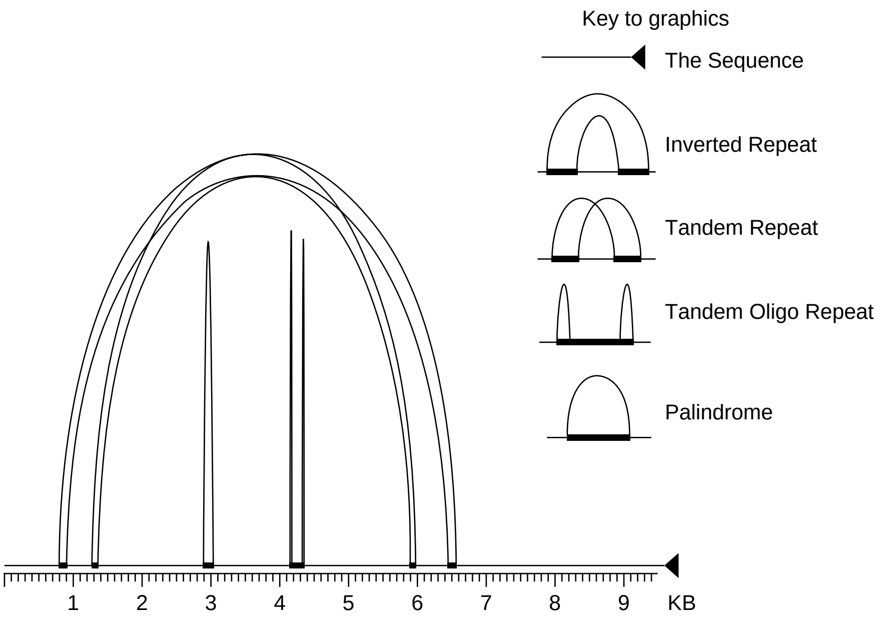


FIGURE 9.2. Miropeats analysis of a genomic region from flax showing inverted repeats, tandem oligo repeats, and a palindrome.

activity is the annotation of the features of the genome. However, because the task is not charged to a single individual or group, how does the genome annotation get accomplished and how do the various features identified by different individuals get integrated? This can be done in a number of ways. One way is the Distributed Annotation System (DAS). This is a client-server system in which a single machine (the client) gathers genome annotation information from multiple distant websites (the reference and annotation servers), collates that information, and displays it to the user in a single view. Little coordination is needed among the various information providers. Further information regarding DAS can be gained at http://biodas.org. A second method involves a group or small sets of groups who generate the annotation and then release the data at regular intervals. If there are multiple annotations of the same region these can be integrated or displayed independently, as is the case for the Human Genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway?org=human). One of the particularly useful functions is the ability to add a personal track to those already established to incorporate individually important features. Finally, for specific gene families, an annotation jamboree could be an appropriate vehicle for annotation. This is where experts in a gene family get together to annotate as many as possible of the representatives of their particular gene

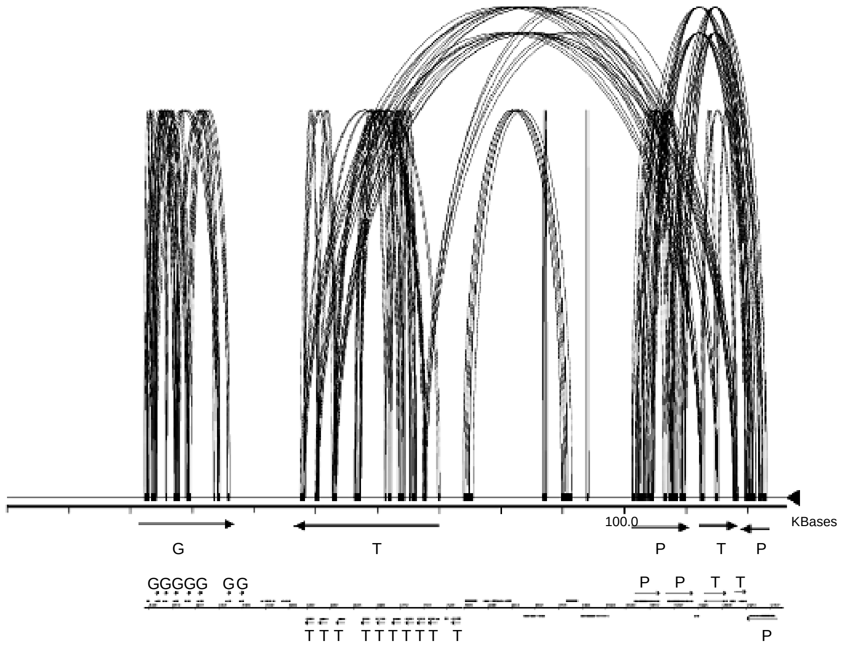


FIGURE 9.3. Miropeats analysis of the *Arabidopsis thaliana* BAC clone F16P2 at a threshold of 150. The organization of genes on BAC F16P2 showing the 3 tandem gene duplications. The glutathione-*S*-transferase and tropinone reductase genes are labeled G and T, respectively. A smaller duplication of pumilio-like protein (P) is also present. Figure 1.4 converted to a single linear read is given below the Miropeats pattern.

family. Clearly, these approaches are not mutually exclusive. The distributed annotation allows experts in various areas to contribute remotely.

The programs that can be used to annotate genomes, and the ways the results are presented, are available at various sites listed in the references for this chapter. Among the most common software tools used are BLAST, FASTA, HMM profiling and motif finding, with Prosite and other pattern/motif combinatorial search tools applied to the various databases listed in Table 9.2. Used alone or in various combinations, they permit the identification and subsequent description of homologous sequences according to various criteria such as sequence, structural, and functional properties. As the available data set that can be searched increases, more identified genes will be available for training the gene finding programs, thereby making their predictions more accurate. The predicted genes can be searched for in the EST databases as well as experimentally by RT-PCR with primers designed from the annotated genomic sequences. In particular, the use of

primers across predicted splice sites would be a direct test of the accuracy of the informatic analyses.

Expression Data

At the highest (most complete) level the genome of an organism should be annotated completely, with all the possible features included in this annotation. Because this cannot be done directly at the outset, it will be built up by using other information to identify interesting and important regions of the genome. Some of these classes of additional data would include the expression data derived from EST sequences, SAGE, and MPSS™ and from comparative genomics.

EST Clustering

As described above, the clustering of ESTs is an important function in understanding the expressed portions of the genomes. One of the sites where this is done is www.tigr.org, where the gene indices are available for 18 different higher plants (*Arabidopsis*, barley , cotton, grape, ice plant, lettuce, *Lotus*, maize, *Medicago truncatula*, *Pinus*, potato, rice, rye, *Sorghum bicolor*, soybean, sunflower, tomato, and wheat). The gene indices at TIGR are the results of the clustering of transcripts into tentative consensus (TC) sequences. These TCs are built with a variety of programs including:

- Megablast (Zhang et al., 2000)
- CAP3
- Paracel TranscriptAssembler™
- DNA-Protein Search program (dps) developed by Dr. Xiaoqiu Huang

An alternative clustering of plant ESTs is available from PlantGDB (http://www.plantgdb.org/). Here the clustering is done with a clustering tool, PaCE, resulting in tentative unigenes (TUGs). The overall data are similar, as can be seen from Table 9.3, although the dates of the relevant assemblies are different, which would explain some of the differences in the numbers in the various classes.

What is clear from both of these sites is that there is still a large number of singletons in addition to the assembled TCs/TUGs. This is especially true for the wheat data, with >70,000 singletons in addition to the 29,000+ TCs/TUGs.

HarvEST (http://harvest.ucr.edu/) is another EST database-viewing software that emphasizes gene function and is oriented ultimately to comparative genomics. This software is downloadable from the website. The EST sequences in HarvEST have also been assembled with CAP3. The fully enabled versions of HarvEST allow the user to examine the actual CAP3

TABLE 9.3. Clustering of EST Data from TIGR and PlantsGDB

Source of data	Plant	Total ESTs	Assembled ESTs	EST Contigs	EST Singlets
Plant GDB	*Zea mays*	206,015	203,358	21,063	19,350
TIGR Gene index	*Zea mays*	192,436	173,826	20,459	15,147
Plant GDB	*Arabidopsis*	178,538	178,464	19,874	29,282
TIGR Gene index	*Arabidopsis*	232,136	216,159	22,485	15,977
Plant GDB	Wheat	415,818	415,642	29,933	77,623
TIGR Gene index	Wheat	415,125	343,891	38,548	71,234

Assembled ESTs are different in the two sets of data. For the TIGR gene index the assembled number is those ESTs included in contigs, whereas for PlantGDB it is all those ESTs used in the contigs and singletons.

sequence alignment and so determine whether and where individual sequences deviate from a consensus sequence.

Therefore, HarvEST and the TIGR Gene indices are similarly assembled with CAP3, whereas the PlantsGDB data are assembled with different software (PaCE).

Finding Genes in Genomic Sequences

The GeneSeqer web service (at the PlantsGDB website) is intended primarily for the purpose of performing the spliced alignment of query sequences (sequences representing transcribed genes, i.e., ESTs, cDNAs, and proteins) with a target sequence (genomic DNA). The input sequences can be determined based on similarity to other sequences (both genomic and transcribed), or one may already possess an uncharacterized sequence one needs to know more about. An exhaustive alignment of "All Plant" ESTs and cDNAs is possible, or a more efficient approach using the Tentative Unique Gene clusters (TUGs) assembled with the PlantGDB contiging method can be made.

Protein Characterization and Comparisons

Again, as with nucleic acid comparisons, a large number of tools are available. One listing with a short description is at http://www.bioinformatik.de/cgi-bin/browse/Catalog/Software/Online_Tools/.

Prosite is a database of protein families and domains (http://us.expasy.org/prosite/) that also makes available an extensive suite of proteomics tools. The database consists of biologically significant sites, patterns, and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs. The underlying basis is that the many different proteins can be grouped into a limited number of families on the basis of similarities in their sequences. Proteins or protein domains belonging to a particular family generally share attributes important for the function of the protein and/or for the maintenance of its three-dimensional structure. This type of analysis of such domains can lead to a protein signature that can be used to assign a newly sequenced protein to a specific family of proteins and thus to formulate hypotheses about its function. Prosite currently contains patterns and profiles specific for more than 1000 protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins. However, the programs, sequence motifs, and domains defined in Prosite and other protein databases have been developed and trained primarily on fungal and animal proteins. Many of these motifs and domains may differ significantly in plants, and so any functional inferences may be substantially improved by retraining the motif descriptions with plant sequences. As the characterization of plant proteins continues it is expected that descriptions for novel plant-specific sequence motifs will be discovered and be useful in predicting the function of unknown plant proteins.

In the same way that there are specialized nucleic acid databases, there are also specialized protein sites. The PlantsP database (http://plantsp.sdsc.edu/), for example, is dedicated to understanding phosphorylation processes in plants, because protein phosphorylation and dephosphorylation are fundamental to cellular regulation. The protein kinase and protein phosphatase families in *Arabidopsis* contain more than 1300 members. The same site has information on the rice protein kinases, where each protein has been assigned to a class, a group, and a family. The assignments are based on the PlantsP Kinase Classification (PPC) (a bottom-up systematic classification based on sequence comparisons using the entire sequence so that sequences that share domains outside of the kinase catalytic domain should cluster together before sequences that only have the catalytic domain in common) based on BLAST searches with an E value cutoff of $1e^{-30}$. These assignments result in five groups:

- Clear (assignment is unambiguous)
- Strong (assignment is highly likely)
- Weak (assignment is to best group, but E values were $>1e^{-50}$)
- Mixed (assignment to more than 1 group possible)
- No assignment for proteins that do not match at an E value $<1e^{-30}$.

COMPARATIVE GENOMICS

The number of phylogenetic programs is continually increasing. One listing contains some 194 phylogeny packages (http://evolution.genetics.washington.edu/phylip/software.html), and no attempt will be made to consider any of the merits or shortcomings of these programs here.

The alignment of genomes across a wide range of species can add substantially to speculation concerning the function of various sequence regions or motifs, and to the identification of genes. These alignments can be done on the basis of sequence similarity, which can be scored on a nucleotide or amino acid basis. An example would be the search for novel regulatory elements with phylogenetic footprinting (Blanchette et al., 2003). In this method the assumption is that functional features evolve more slowly than nonfunctional features. Where short motifs have been well conserved across evolution, two different challenges arise. If the sequences are closely related, it may be relatively easy to obtain a good multiple alignment. However, when sequence divergence occurs, distinguishing between chance observation versus selection pressure requires quantitative analysis. A statistical test that uses known phylogenetic relationships between the species has been developed to determine whether the region is likely to be under selective pressure (Blanchette et al., 2003). In one case, for closely related organisms, a parsimony score is calculated and a *P* value associated with this is reported, using a null model for neutral evolution. In the second case, when the promoter regions are quite diverged, a program called FootPrinter, which identifies conserved motifs from a set of unaligned sequences, is allied to the phylogenetic information to evaluate a motif's conservation. Results obtained with a large genomic region in 13 vertebrates is shown on the genome browser at the University of California, Santa Cruz (http://hgwdev.cse.ucsc.edu/zoo.html). The conserved regions that are identified from these types of comparisons can then be subjected to experimental verification. One of the specific advantages of using these comparative approaches is the identification of important regions that lie far away from the gene in whose control they are involved. A region that is approximately 60 kb 5′ to the mouse Gdf6 gene (required for the development of several skeletal joints) was shown to be critical for proximal limb joint expression. This was identified with BAC transgene constructs and analysis of transgenic mice. The identified region contained several evolutionarily conserved noncoding sequences that are likely to be the core components of a limb joint regulatory element. However, there is evidence to suggest that additional regulation elements lie outside of the ~250-kb region that has been extensively characterized with the BAC transgenes. MultiPipMaker (an informatic tool developed by Webb Miller, http://nog.cse.psu.edu/pipmaker/) was used to compare those regions identified across a range of vertebrates, resulting in the identification of a rodent-specific syntenic break

close to Gdf6, thereby making it unlikely that the additional regulatory elements could be identified by using sequence comparisons across a wide range of mammals (Mortlock et al., 2003).

As described in Chapter 1, syntenic relationships can be identified among the flowering plants and beyond. It still awaits verification whether the same type of long-range regulatory signals can be identified in plants as have begun to be identified in animals. However, as the number of genomic sequences increases, and the level of genomic sequence in a wider range of plants becomes available, this type of analysis will become more feasible. One of the advantages in animals is the availability of larger numbers of "complete" genomic sequences. The proposals for plant genomic sequences that are focused on the expressed or "genic" regions may fail to include those regions responsible for any long-range regulation, so a case can be made for including the "unimportant" repetitive regions in the sequencing of complex plant genomes.

The genomic context of genes is also an essential piece of information to help distinguish between paralogs and orthologs. The alignments that are being developed and are available from the specific databases, such as Gramene and MaizedB, are important resources in this task. The availability of browsers that support a custom annotation track will allow individual users to build their own annotation structures, and these eventually will substantially contribute to the distributed annotation effort.

Interpretation of Data

The analysis of sequence data usually results in an extensive output that must be scrutinized for its applicability to the question being asked. A BLAST search of nucleotide data will generate many matches at differing E values (the expected number of chance alignments), and a decision must be made as to the relevance of each. In other words, what is the appropriate cutoff point for considering a match to be significant?

The probability of the match being made by chance depends on the size of the query sequence and the size of the target database. If the aim is to find other plant sequences that are related to the one under consideration, then a search of that subset of the sequence databases would be most appropriate. If the aim is to find any related sequence, then a more inclusive set of sequence databases, or the whole nonredundant sequence database, would be most appropriate.

Nucleic Acid Versus Protein Homology Searches

The difference in the search for nucleic acid homology versus protein homology also impacts the type of comparison being done. The data in Figure 9.4

BLAST 2 SEQUENCES RESULTS VERSION BLASTN 2.2.5 One ubiquitin coding unit from flax polyubiquitin gene and the Cucumis melo polyubiquitin mRNA, partial cds Accession number AF436850.

Score = 162 bits (84), Expect = $5e^{-37}$
Identities = 170/213 (79%)
Strand = Plus / Plus

```
Query: 134 aagaccctgacgggtaaaaccatcaccctcgaggtcgaatcctctgataccatcgataac 193
           ||||||| ||| || || ||||||||| ||||| ||   |   || |||||| |||
Sbjct:  16 aagacccttgactggcaagaccatcaccctggaggtggagagcagcgacaccatcgacaac 75

Query: 194 gtcaaggcaaagatccaggacaaggagggaattcccccagaccaacaaagactcatcttt 253
           || || || ||||||||||||||| ||||| || || ||||| || ||  || ||||||
Sbjct:  76 gtgaaagctaagatccaggacaaggaaggaatccctccggaccagcagaggctgatcttt 135

Query: 254 gctggtaagcaattagaggacggccgtacccttgccgattacaacatccagaaggagtcc 313
           || || ||||  |  | || |||| | || |  | ||||||||||||||| ||||||| ||
Sbjct: 136 gccgggaagcagctggaagacggcaggactttggccgattacaacattcagaaggaatca 195

Query: 314 accctccaccttgtgttgcgtctccgtggtggt 346
           ||||| |||||||||| || | ||||||||||||
Sbjct: 196 acccttcaccttgtgctgaggctccgtggtggt 228
```

Part of the output from blastx of the same flax fragment highlighting the matches found for polyubiquitin from *Cucumis melo*

Score = 149 bits (376), Expect = $1e^{-35}$
Identities = 75/76 (98%), Positives = 76/76 (100%)
Frame = +1

```
Query: 1   MQIFLKTLTGKTITLEVESSDTIDNVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLADYN 180
           MQIF-KTLTGKTITLEVESSDTIDNVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLADYN
Sbjct:  40 MQIFVKTLTGKTITLEVESSDTIDNVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLADYN 99

Query: 181 IQKESTLHLVLRLRGG 228
           IQKESTLHLVLRLRGG
Sbjct: 100 IQKESTLHLVLRLRGG 115
```

Score = 80.9 bits (198), Expect = $5e^{-15}$
Identities = 39/39 (100%), Positives = 39/39 (100%)
Frame = +1

```
Query: 112 PDQQRLIFAGKQLEDGRTLADYNIQKESTLHLVLRLRGG 228
           PDQQRLIFAGKQLEDGRTLADYNIQKESTLHLVLRLRGG
Sbjct:  1  PDQQRLIFAGKQLEDGRTLADYNIQKESTLHLVLRLRGG 39
```

Score = 55.1 bits (131), Expect = $3e^{-7}$
Identities = 27/28 (96%), Positives = 28/28 (100%)
Frame = +1

```
Query: 1   MQIFLKTLTGKTITLEVESSDTIDNVKA 84
           MQIF+KTLTGKTITLEVESSDTIDNVKA
Sbjct: 116 MQIFVKTLTGKTITLEVESSDTIDNVKA 143
```

FIGURE 9.4. Blast 2 sequences results version blastn 2.2.5 and part of the output from blastx of the same flax fragment highlighting the matches found for polyubiquitin from *Cucumis melo.* One ubiquitin coding unit from flax polyubiquitin gene was used for the searches. The same flax unit was used with Blast align against the *C. melo* polyubiquitin mRNA, partial cds accession number AF436850. Results for nucleic acid and protein comparisons are shown.

are for one of the repeat units from the flax polyubiquitin gene previously described (Agarwal and Cullis, 1991). Here the search was done with both blastn and blastx. The output shows that although many of the related plant ubiquitin genes are only homologous at 80% of the nucleotides, the proteins are identical. The *Cucumis melo* polyubiquitin mRNA is only 79% homologous to the flax fragment and was not represented in the list of nucleic acid matches resulting from the blastn analysis, whereas it was present in the blastx results. Here the E value varies again with the length of the homology, so that a match with all 75 of 76 amino acids identical (but 76 out of 76 positive) gives an E value of $1e^{-35}$, whereas 39 out of 39 identities gives an E value of $5e^{-15}$ and 28 out of 29 identical (but 29 out of 29 positive) gives an E value of $3e^{-7}$.

Similarly, the data from gene prediction programs must be scrutinized carefully. The data from a flax sequence (the same one used for the Miropeats analysis in Figure 9.2) were used with the GeneSeqer program at www.PlantsGDB.org for the prediction of possible transcripts from this region. This sequence was from a region of the genome that has an insertion in some flax varieties. Three possible transcripts were identified, none of which was very long (174 bases, 86 bases, and 170 bases, respectively). The possible transcript of 170 bp crossed the boundary between one end of the insertion sequence and the surrounding region. However, when inspected only three of the matched bases were included in the coding sequence of the matched EST, whereas the remaining corresponding nucleotides were between an A-rich region in the genomic sequence and the polyA tail of the EST. Again this reinforces the necessity that informatic analysis conclusions, purely by themselves, not be taken as correct and conclusions not be drawn on this basis alone.

SUMMARY

The enormous amounts of data being generated by high-throughput genomics have resulted in the need for developing expertise in the analysis and storage of the data. The possible scale of the data that can range from the description of a plant phenotype, through a complete genome sequence, to the microarray analysis of all of the possible open reading frames under various perturbations raises a number of new problems. How are the data to be stored and interconnected, and for how long? As projects come to their conclusion, what data are archived and remain accessible to the scientific community, and what data disappear? The decisions about the longevity and permanent archiving of data must be taken before the data are lost (or become irretrievable). Organism-specific and more general databases have been and are being developed for storage and manipulation of all of these data. The websites associated with many of these databases also provide

access to the tools for the analysis of the data. A large number of informatics tools are currently available, and it is becoming increasingly difficult to decide which particular one is most appropriate to use. Because the development of many of these programs still continues, with new programs also being introduced, it is important that those that are available are supported in some way, or at least hardened so that they are fully functional. As always, it is important to remember that just because the informatic analysis provides a particular answer, the robustness of that conclusion must always be checked.

This chapter does not deal with the needs of microarray analysis or proteomics, which have been covered in relevant sections elsewhere. However, many of the sites referenced in this chapter also contain information and programs relevant to these techniques; for example, the TIGR site has a great deal of information about microarray experimentation and analysis.

References

Agarwal, M. L., and C. A. Cullis (1991) The ubiquitin genes in flax. *Gene* **99**, 69–75.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin. and G. Sherlock (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29.

Blanchette, M., M. Tompa, and D. Haussler (2003) Advances in Genome Biology and Technology, Marco Island. Abstract p. 11.

Mortlock, D., C. Guenther, and D. Kingsley (2003) Advances in Genome Biology and Technology, Marco Island. Abstract p. 85.

Zhang, Z., S. Schwartz, L. Wagner, and W. Miller (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214.

Web Resources

http://www.bioinformatik.de/cgi-bin/browse/Catalog/Software/Online_Tools/

http://us.expasy.org/prosite/

http://evolution.genetics.washington.edu/phylip/software.html

http://hgwdev.cse.ucsc.edu/zoo.html

http://nog.cse.psu.edu/pipmaker/

http://www.PlantsGDB.org

http://www.geneontology.org/)

http://www.genome.ou.edu/SeqStrategy.html

http://www.genome.gov/10001812

http://www.ncbi.nlm.nih.gov/BLAST/)

http://www.ncbi.nlm.nih.gov/BLAST/blast_overview.html
http://www.genome.ou.edu/miropeats.html
http://biodas.org
http://genome.ucsc.edu/cgi-bin/hgGateway?org=human
http://www.tigr.org

CHAPTER

10

Bioethical Concerns and the Future of Plant Genomics

Overview

The increased understanding of the structure and function of plant genomes and the ability to manipulate those functions have far reaching implications; for example, the application of this knowledge for the improvement of crop plants is an important activity to ensure the world food supply. Many of the basic discoveries concerning the function of particular genes, and more specifically the characteristics conferred on an individual by a combination of specific alleles, can be included in conventional breeding techniques. However, the ability to transform plants and introduce genes from any source has opened a much more elaborate toolbox for considering and applying novel approaches to plant improvement. The same technology that has been used to confirm many of the conclusions about the function of particular genes, namely the development of transgenic plants, has also been extensively used in the development of new crop varieties, particularly those in which genes for herbicide and insect resistances have been introduced.

Traditional plant breeding, which involves the development of new varieties by rounds of intercrossing and selection, has been very successful in improving the performance of crop plants. However, even if advantage is taken of using various parts of the world to complete multiple generations of breeding in a single calendar year, new varieties still take many years to develop and release. Therefore, both the potential of transgenic methods to circumvent many generations of backcrossing and the use of marker-assisted selection to improve the efficiency of the identification of the desired

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

combination of genes have impacted plant breeding. In addition to the direct scientific impacts, the development of the molecular techniques has also siphoned support from traditional approaches to plant breeding, both in the actual breeding and in training in this area. Thus public funding support has been redirected away from conventional plant breeding and toward molecular genetic approaches (Knight, 2003). The situation becomes more involved when we consider that the trend away from public plant breeding efforts to commercial sector breeding programs has also altered the spectrum of crops and traits involved. However, windfalls of data can occur when these commercial programs are discontinued, such as the release into the public sector of the wheat EST data from Dupont when it discontinued its wheat breeding program.

Because commercial breeding efforts are geared toward generating profits, the development of varieties that are suitable for growth in some of the challenging environments in developing countries will not necessarily be commercially attractive. Therefore, it is important that efforts are made to transfer the technical expertise to these countries, along with the infrastructure to enable the application of genomics discoveries to food security issues of the developing world.

Public understanding of the science and technology underlying the production of transgenic crops is crucial to the future of this technology. Without the appropriate knowledge it will be impossible for the public to make objective decisions regarding the pros and cons of genetically modified (GM) food production and consumption. This is clearly a challenge where science literacy is relatively poor and a number of different factions are taking very disparate stands on both the safety and benefits of this technology.

What is Being Grown?

The two most common types of genes currently being used in transgenic crops for commercial production are those that confer herbicide resistance and insect resistance (Bt genes). Soybean, corn, cotton, and canola are the four major GM crops currently being grown. The amount of land devoted to growing GM crops has continued to increase (Figure 10.1). However, although GM crops were grown in 16 countries in 2002, just 4 countries (USA 66%, Argentina 23%, Canada 6%, and China 2.1%) accounted for 99% of the total global plantings (James, 2002).

The major issues relating to genetically modified organisms (GMOs) fall into four general areas:

1. Safety, including:
    a. The safety of the product itself to the end user
    b. The risks of releasing the GMO into the environment

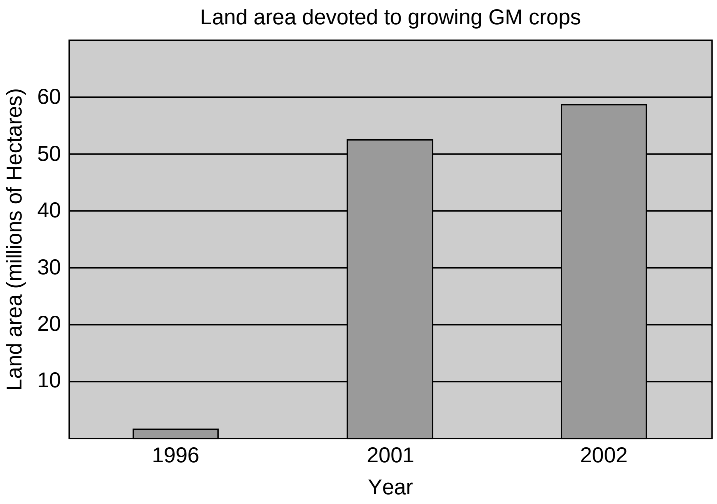


FIGURE 10.1. Land area devoted to growing GM crops (Data from James, 2002).

2. Trade issues: The acceptance or rejection of GM foods by a nation impacts import regulations.
3. Developing regulations for the testing and release of GMOs.
4. Who is, and/or should be, deriving benefit from GM products—the consumers, producers, farmers, and/or technology developers?

Safety Issues

Safety concerns include the safety of GM foods themselves and the safety of releasing living modified organisms (LMOs) into the environment. These two issues clearly impact the adoption of the cultivation of transgenic plants as well as trade involving such commodities.

The introduction of any safety legislation in a country has usually followed a major accident or incident. Such regulations are therefore reactive. In the case of modern biotechnology the system of regulation has been proactive right from the outset. In the USA the development of protocols for the testing and release of GM plants occurred before any release of transgenic plants. This proactive generation of regulations continues despite the lack of documented cases of harm resulting directly from the use of recombinant techniques either in the research environment or in commercial applications. One highly publicized study involved the feeding of monarch caterpillars on *Asclepias* spp. (milkweed) leaves heavily dusted with Bt-corn pollen, with

the reported result that the caterpillars suffered, thereby triggering concerns about Bt crops (Losey et al., 1999). Subsequent studies demonstrated that, under actual field conditions, the likelihood of monarch caterpillars being exposed to damaging levels of Bt-corn pollen was low (Wraight et al., 2000; Zangerl et al., 2001; Stanley-Horn et al., 2001; Shelton and Sears, 2001; see also http://www.ars.usda.gov/sites/monarch/index.html for a full discussion of the history and conclusions of this controversy).

In light of this, the question arises of how and why biotechnology is so fundamentally different from any other technology that it needs to be regulated in such a different way. Is it possible that any of the innovations that so fundamentally modified our way of life during the twentieth century would have happened had a full risk evaluation been required? The proactive approach implies a concern that there is something to fear from the new technology, a concern that is particularly pervasive in Europe.

Trade Issues

Why are issues about the movement of GM products important? The major growers of GM crops are also major exporters (Table 10.1) but only comprise a very small fraction of the nations that consume these crops. Therefore, the worldwide acceptance is important in the propagation of GM crops and their adoption. In this respect, the USA, joined by Argentina, Canada, and Egypt, has recently (2003) filed a complaint to the World Trade Organization (WTO) challenging the European Union's (EU) *de facto* moratorium on the commercialization of GM products, arguing that the EU's moratorium on the approval and commercialization of GM products is an illegal trade barrier under WTO rules. The European Commission countered by calling the move legally unwarranted, economically unfounded, and politically unhelpful and arguing that the EU's regulatory system for GMO authorization is in line with WTO rules.

TABLE 10.1. Production and Trade of GM Agri-Food Products, 2000

Crop	Number of producing countries	Percentage of global exports from GM producers	Number of importing countries
Maize/corn	8	85	168
Soybeans	6	88	114
Canola	2	50	68

From Pardey and Ko (2003).

Public Perception

Even though there are no documented cases of harm resulting from the use of this technology (people around the world have been eating biotech food for years) the public perception of GMOs and their products has not always been positive. For example, the consumer perception of risk associated with eating of GM soybeans is less when the consumer is given information regarding consumer benefits than when no informa-tion concerning benefits is provided (Brown and Ping, 2003). Therefore, consumer perception of risk associated with the consumption of GM-derived foods is affected by the direct associated benefits detailed to the consumer. The response in Western Europe to the new technology cannot simply be alleviated through assertions by scientists that there is negligible risk. Nor can the safety of GM foods be based solely on scientific risk as-sessments unless there is a scientific consensus that the impact on the environment is negligible. Therefore, the approach such as that at http://www.ars.usda.gov/sites/monarch/index.html "Butterflies and Bt corn—Allowing science to guide decisions," may not be very effective in swaying public opinion.

There must be some tangible benefit to the consumer, not just the producer or supplier, to make the introduction of the new technology reasonable. In the developing world, the benefits to the small farmer from the increase in economic activity and the improvement in the living standards for a large proportion of the population may be a sufficient perceived benefit to the consumers to promote the adoption of the new technologies.

However, this circumstance, in itself, raises additional problems with respect to the framework and resources in place in many developing countries to permit the growth and distribution of GMOs.

Case Study

Lesotho, Malawi, Mozambique, Swaziland, Zambia, and Zimbabwe raised concerns about damaging their future trade with the EU if they accepted food aid that included genetically modified corn, even though a significant portion of the population was starving. The primary concern was that farmers could plant the US-donated corn kernels and if those plants cross-pollinated with the local corn varieties the EU could bar their corn or corn-fed animal exports. The situation was exacerbated in that the countries involved did not have adequate protocols to manage GM crops. This problem could have been negated by the supply of ground corn, removing the possibility of the GM material being planted, but USAID donations would not cover the cost of milling.

The UNEP International Guidelines and the Cartagena Protocol states:

The Conference of the Parties to the Convention on Biological Diversity adopted a supplementary agreement to the Convention known as the Cartagena Protocol on Biosafety on 29 January 2000. The Protocol seeks to protect biological diversity from the potential risks posed by living modified organisms resulting from modern biotechnology. It establishes an advance informed agreement (AIA) procedure for ensuring that countries are provided with the information necessary to make informed decisions before agreeing to the import of such organisms into their territory. The Protocol contains reference to a precautionary approach and reaffirms the precaution language in Principle 15 of the Rio Declaration on Environment and Development. The Protocol also establishes a Biosafety Clearing-House to facilitate the exchange of information on living modified organisms and to assist countries in the implementation of the Protocol

Cartagena Protocol on Biosafety http://www.biodiv.org/biosafety/

and so mandates that the public is to be informed and educated about biosafety. However, the reaction against the technology in Europe has greater impacts on the public image elsewhere than can be generated from a reasoned argument for the safe use of the technology. In addition, many parts of the developing world have not put in place the framework for dealing with the issues of GM crops and their products.

It is relatively simple for countries to set up a legal framework. In fact, many have done so, and in this regard the Zambian government has developed a National Biosafety and Biotechnology Strategy Plan for 2003 to 2007 in response to debate over GM foods. The five-year plan will take care of the unwarranted proliferation of GMOs in the country and set a local design for the development of biosafety regulations to protect the country's unique biodiversity.

Although the legal framework can easily be put in place, it is a much longer process to provide the scientific infrastructure to produce new food varieties. Again using the Zambian example, it is estimated that the implementation of this plan will require an estimated $18 million for the development of human resources and infrastructure alone to deal with the GMO issues.

Therefore, according to the Cartagena Protocol, countries must institute processes to ensure that LMOs are used safely. These needs are many and complex and include:

- A versatile and flexible system for receiving notifications about the intended introduction of a LMO that can handle both contained and released organisms
- Ensuring completeness of the notification so that all the necessary information is provided
- Scientific oversight of the risk assessment whether it is performed in the country or from data provided

- Systems that allow public and stakeholder comment to be considered before any decision being made
- The inclusion of modifications as a result of public comment
- The release of the decision to the party concerned and the public

Most important of all, this structure must be in place before the first decision is, or can be, made. As is required in item 3, there must be appropriate scientific resources to evaluate the data, and these resources are currently unavailable in many of the developing countries.

As the distribution of GM crops increases, there will be many opportunities for countries to cooperate in the decisions concerning the release of such crops. Because many of the parameters would be similar, it may not be necessary for neighboring countries to repeat the complete process for a product that has already been approved by a neighboring country; for example, would Kenya and Uganda need to independently repeat all the steps to approve a product once it has been approved in one of these two countries?

At present, there is a growing interest in biosecurity because of major international developments such as the globalization of the world economy, rapid increase in communications, transport, and trade, technological progress, and an increased awareness of biological diversity and environmental issues. The international regulatory institutions that impinge on world food trade issues are shown in Table 10.2.

The international regulations covering movement of GMO (or LMOs) and the number of countries that agree to these regulations are also shown in Table 10.2. The Cartagena Protocol required a minimum of 50 signatories for formal adoption, a number that has now been reached.

The adoption of biotechnology and the introduction of GM foods into the international marketplace have further complicated an already difficult area of trade policy. Diverging national regulations and public perceptions have impacts on the trade in these products. The recent WTO agreement on agriculture has not yet been fully implemented, and many of the issues still outstanding are highly contentious. Because there is little goodwill in the policy community that can be directed to resolving the growing trade irritants caused by GM foods, the private sector may need to change how it introduces and markets the new products of biotechnology to maintain market access (Pardey and Ko, 2003).

In Africa, according to their ministers of agriculture, farmers urgently need appropriate modern biotechnology packages to help improve the declining production of major food and cash crops. However, if African and other developing nations are not to be marginalized further, but rather benefit from biotechnology, they must have initiatives that include improving their expert capacity for science and technology by producing a core of highly skilled scientists or experts. In addition, the next generation of trans-

TABLE 10.2. International Regulatory Institutions

Institution	Members	Coverage
Food and Agriculture Organization of the United Nations (FAO)	184	Food security programs
World Health Organization (WHO)	191	Health science and policy
International Plant Protection Convention (IPPC)	107	Pests and pathogens (crops)
International Epizootics Organization (OIE)	155	Pests and pathogens (animals)
Codex Alimentarius (Codex)	165	Food standards and labels
World Trade Organization (WTO)	139	Trade rules for all goods;
Dispute Settlement Mechanism		
Organization for Economic Cooperation and Development (OECD)	29	Harmonize standards and policies
Regional Initiatives	Various	Harmonize science or processes
Cartagena BioSafety Protocol (BSP)	Minimum 50	Transboundary movements of living modified organisms

From Pardey and Ko (2003).

genic crops should have definite consumer advantages, such as better nutritional value in a culturally acceptable form, to overcome the possibility that in developing countries the emphasis on the risks and dangers of biotechnology will, one day down the road, overshadow the positive benefits of the technology.

Public Policy Issues

It is clear that scientists need to be involved in the public debate over the merits of GM foods and crops, and many individuals and organizations have generated material that is useful in engaging the public in such a debate (http://ucbiotech.org/; http://www.biotech.iastate.edu/Educational_resources.html; http://www.biotech.wisc.edu/; http://www.nal.usda.gov/bic/Education_res/). However, lack of harm, by itself, is insufficient justification for the introduction of such foods and crops. The following factors should be addressed in any debate on the merits and risks associated with these products (Baline et al., 2002):

- Consumers do not reject GE outright but focus on specific applications of the technology.
- Media coverage and activist opposition are associated with more negative perceptions.
- Overall, support for the technology tends to be dependent on perceptions of benefits.
- Consumers view risk very differently from scientists.
- The amount of media coverage, not the content, creates the negative perceptions.
- Consumer awareness of GM foods has increased, but understanding remains low.
- Consumer acceptance of biotechnology tends to be conditional and dependent on perceptions of benefits and moral acceptability.
- The single most important determinant of gain or loss of trust in a source is whether the information is subsequently proven right or wrong and whether the source is subsequently demonstrated to be unbiased.

New Avenues for GM Products

Many of the problems in the developing world relate to stress tolerances, which are complex traits. These will be much more difficult to manipulate with only the introduction of single transgenes. Therefore, the strategies for altering these properties may need to be developed by using model systems to identify the important components. With the improvements in transformation technology and understanding of precise gene regulation, it may become possible to introduce a suite of genes to generate a new variety. If these genes are from related species and could conceivably be introduced through conventional breeding schemes (albeit slowly and inefficiently) will the new varieties be considered as GMOs? If so, will the restrictions be less or more than for those containing novel nucleic acid sequences? A proposed categorization of transgenic organisms (Nielsen, 2003) may help in the debate concerning the regulations covering various classes of transgenics. In the same vein, the suggestion for defining genomics-guided transgenics (Strauss, 2003) looks to differentiate between different types of transgenic organisms and the risks associated with each.

Summary

Evidence from industrial and developing countries show that current GM crops, in conjunction with conventional agricultural practices, can contribute

to a cost-effective, sustainable, productive, and sufficiently safe form of agriculture (Conner et al., 2003). The likely impacts that have been identified with GM crops are similar to those of new cultivars that have been derived from traditional breeding. Therefore, the appropriate yardstick with which to evaluate the risks of GM crops may be the risks that could arise from any new variety.

As the understanding of the processes that are used to develop GM crops increases and the information concerning the safety and efficacy of GM crops becomes more widely available, it is likely that the public acceptance of these crops will also increase. In addition, a wider range of modifications in the available GM material, especially with respect to direct customer benefit, is an important aspect for the wider acceptance of GM foods. However, as many of the important traits that need to be modified, especially of crops in the developing world, are more complex and less well understood than those products already in production, substantial delay in the introduction of such material is likely to occur.

References

Baline, K., S. Kamaldeen, and D. Powell (2002) Public perceptions of biotechnology. *J. Food Sci.* **67**, 3200–3208.

Brown, J. L., and Y. Ping (2003) Consumer perception of risk associated with eating genetically engineered soybeans is less in the presence of a perceived consumer benefit. J. Am. Dietetic Assoc. **103**, 208–214.

Cartagena Protocol on Biosafety at http://www.biodiv.org/biosafety/).

Conner, A., T. Glare, and J.-P. Nap (2003) The release of genetically modified crops into the environment: Overview of ecological risk assessment. *Plant J.* **33**, 1–18.

James, C. (2002) Global Review of Commercialized Transgenic Crops: 2001(Feature: Bt Cotton). ISAAA Briefs No. 26. Ithaca, NY: ISAAA.

Kinderlerer J. Regulation of Biotechnology: Needs and Burdens for Developing Countries http://www.unep.ch/biosafety/BTregulationJK.pdf.

Losey, J. E., L. S. Rayor, and M. E. Carter (1999). Transgenic pollen harms monarch larvae. *Nature* **399**, 214.

Nielsen, K. M. (2003) Transgenic organisms—Time for conceptual diversification? *Nat. Biotechnol.* **21**, 227–228.

Pardey P. G. and Ko (eds) (2003) *Biotechnology and Genetic Resource Policies*. Washington, DC: International Food Policy Research Institute.

Shelton, A. M., and M. K. Sears (2001) The monarch butterfly controversy: Scientific interpretations of a phenomenon. *Plant J.* **27**, 483–488.

Stanley-Horn, D. E., G. P. Dively, R. L. Hellmich, H. R. Mattila, M. K. Sears, R. Rose, L. C. H. Jesse, J. E. Losey, J. J. Obrycki, and L. C. Lewis (2001). Assessing the impact of Cry1Ab-expressing corn pollen on monarch butterfly larvae in field studies. *Proc. Natl. Acad. Sci. USA* **98,** 11931–11936.

Strauss, S. H. (2003) Genomics, genetic engineering, and domestication of crops. *Science* **300**, 61–62.

Wraight, C. L., A. R. Zangerl, M. J. Carroll, and M. R. Berenbaum (2000). Absence of toxicity of *Bacillus thuringiensis* pollen to black swallowtails under field conditions. *Proc. Natl. Acad. Sci. USA* **97**, 7700–7703.

Zangerl, A. R., D. McKenna, C. L. Wraight, M. Carroll, P. Ficarello, R. Warner, and M. R. Berenbaum (2001). Effects of exposure to event 176 *Bacillus thuringiensis* corn pollen on monarch and black swallowtail caterpillars under field conditions. *Proc. Natl. Acad. Sci. USA* **98**, 11908–11912.

AFTERWORD

The main point to make after all the descriptions of technology contained here is not to lose sight of the main goal, that is, to understand the biology. As Sydney Brenner so succinctly put it at the Plant and Animal Genome Meeting in January 2003:

> *Forget about the genome and get on with the biology.*

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1 Copyright © 2004 John Wiley & Sons, Inc.

Index

Abiotic interactions, of plants, 141–144
Ac transposon, 82, 85
Adh1-F region, 11
Africa, genetically modified organisms in, 193–196
Agrobacterium tumefaciens, 28, 93
Aligned gene maps, 17, 18
Alleles, 147
Allopolyploidy, genome size and, 13–14
Alternative splicing, 112
Amino acid, sequence from proteomics experiments, 121, 122
Amplified fragment length polymorphisms (AFLPs), 149, 151
  techniques using, 2–3
Annotation, 77–78, 79
  informatics for, 174–179
AP1/EREBP gene family, 97
Arabidopsis, 54, 55, 57, 70, 95, 119, 120
  annotated genome of, 78
  centromeres of, 12–13, 19
  complete genome sequencing of, x–xi, xii, xiii, 3, 47, 49, 53
  genome size of, 2
  genome variation in, 7–10, 15
  GUS activity in, 82, 84–85
  Miropeats analysis of, 174, 177, 178
  molecular maps for, 149
  mycorrhizal fungi and, 139
  pathogen resistance mechanisms of, 135, 136
  phenotypic effects of genes in, 147
  sequencing strategies for, 64–65
  T-DNA insertion in, 81
  transcription factor families in, 96, 97–98
Arabidopsis Biological Resource Center (ABRC), 81
Arabidopsis gene DDM1, 102
Arabidopsis Knockout Facility, 81
Arbuscular mycorrhizal fungi, 138–139
ARF-Aux/IAA gene family, 98
Automated sequencing strategies, 38
Autopolyploidy, genome size and, 13–14
Avirulence (Avr) genes, 133–135
Avr-Pita protein, 137

BAC clones, 12, 27
  in automated sequencing strategies, 38
  data processing and, 63
  in MTP sequencing, 59, 60, 61
  physical mapping and, 51–53
BAC contigs, genome fractionation via enriched, 58
BAC end sequencing, 59–61
BAC F16P2, 8, 9
Backcross breeding schemes, in marker-assisted selection, 160–162
BAC libraries, 28–32
  generation of, 28–30
  in QTL identification, 155
  utilization of, 30–32

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1

Copyright © 2004 John Wiley & Sons, Inc.

Bacteria, nitrogen-fixing, 138–139
Bacterial artificial chromosomes (BACs). *See also* BAC entries
automated sequencing of, 38
large-insert vectors and, 27
libraries of, 28–32
Bacteriophage λ, large-insert vectors and, 27
"Bait" plasmids, 126, 127
Banana, BAC libraries from, 32
Basic Local Alignment Search Tool (BLAST®), 173–174, 175, 176, 177, 178
data interpretation using, 183–185
Beijing Genomics Institute, 48, 49
Beneficiaries, of genetically modified organisms, 191
Benzothiodiazole-inducible Pr-1a system, 101
Bermuda standard, 171
bHLH/MYC gene family, 97
Bidirectional promoter, 93, 94
Binary BAC (BIBAC) vector, 28
Bioethics, of plant genomics, 189–198
Bioinformatics, 167–186. *See also* Informatics tools; Software
Biological databases. *See* Databases
Biotechnology, safety of, 191–192, 193–196. *See also* Bt crops; Technology
Biotic interactions, of plants, 132–141
BP reaction, 25
Brachypodium distachyon, genome size of, 3
Brassica oleraceae, as sequencing target, 65
Breeding methods, in marker-assisted selection, 160–164
Brenner, Sydney, 201
Bt crops, safety of, 191–192, 193
bZIP gene family, 97

Caenorhabditis elegans, 54
CAP3 program, 179–180
CAP4 program, 43
Cartagena Protocol, on genetically modified organisms, 193–196
CC-NBS-LRR gene family, 134, 136
cDNA libraries, 33, 34–35
Celera, 63–64
CEL I mismatch cleavage enzyme, 84
Cells, process control in, 107–108
Cellular homeostasis, in plants, 141
Cell walls, synthesis of, xii
CEN2 gene, sequence features of, 12
CEN4 gene, sequence features of, 12
Centromeric DNA, 12–13
Chemical regulation, of gene expression, 99, 100–101
"Chips." *See* DNA microarrays
Chromatin, 90, 91
structural effects of, 99–102
Chromosomes
evolutionary relationships and, 16–19
RFLPs and, 150
variation in plant, 3–7
cis-acting sequences, 95
Class I and class II transposons, 10, 12
Clonase™, 26, 27
Clones. *See also* BAC clones; Cloning
of complementary DNA, 32–33
number necessary to screen (table), 27
Cloning, 70. *See also* Clones
map-based, 48
Cloning systems, 24–38
CLUSTAL W software, 43
Clustering protocols, for sequence data processing, 63
Coiled coil (CC) genes, 134
Comparative genomics, 182–183
Complementary DNA (cDNA). *See also* cDNA libraries; Concatenated cDNA sequencing (CCS)
in cells, 107–108
in cloning, 32–33
in differential display technology, 36
in gene expression analysis, 110, 112, 113–114
in generating ESTs, 71–73
generating full-length, 73–75
in genome annotation, 78
in microarrays, 37
MPSS and, 117, 118
SAGE and, 114–116
sequencing of, 39, 40, 75–77

sequencing software for, 43
in transgenics, 128
in yeast two-hybrid systems, 126, 127
Complex traits, 147–164
Concatenated cDNA sequencing (CCS), 77
Concentration times time (Cot) parameter, 55–56. *See also* Cot entries
Consumers, attitude toward GMOs, 196–197
"Contamination," organellar, 61
Contigs, in generating ESTs, 72. *See also* BAC contigs; Fingerprint contigs (FPC) software; Tentative contigs (TCs)
Control
of gene expression, 89–103
translational, 103
Conventional breeding methods, in marker-assisted selection, 163–164
Cooperative State Research Education and Extension Services (CSREES), 132
Copper-inducible ACEI system, 101
Cot curves, 55–56
Cot fractionation, 55–56
Cot fractionation-based libraries, 62
Crops, genetic improvement of, xiii–xiv. *See also* Bt crops; Plants; Transgenic crops
Crossbreeding, bioethics of, 189–190
Cucumis melo, BLAST® sequences for, 183–185
Cy3 fluorochrome, 111
Cy5 fluorochrome, 111
Cypripedium calceolus, DNA content of, 2–3
Cytosine, methylated, 57, 62

Data
analyzing genomic, 23–43, 167
expression, 179
gene identification from sequence, 70–73
interpreting, 183–185
static and dynamic, 169
Database-independent software, 4
Databases, 167, 168–170
for *Arabidopsis* gene insertions, 81
creation of, 39–42
critique of, 41–42
developing, 168–169
from NCBI (table), 176
protein, 181
in proteomics experiments, 123
specialized, 170
Data management, 168
Data mining, 168
Data processing, of shotgun sequence data, 63
Data storage, 169–1701
Delayed nodulation mutants, of legumes, 139
Deletion lines, 150
Dexamethasone-inducible GR fusions, 100
Dexamethasone-inducible GVG system, 100
Dex-inducible and tet-repressible TGV system, 101
DICER enzyme complex, 86, 87
Dicots, intron-mediated enhancement in, 95
Differential display technology, 36–37
Diploid plants, 1
Diseases, 131
plant resistance to, 132–138
Dispersed repetitive sequences, DNA variation via, 10–11
Distributed Annotation System (DAS), 177
DNA, 129. *See also* Recombinant DNA; T-DNA insertion
annotated sequences of, 79
in cells, 107–108
in centromeres, 12–13
in chromatin, 102
controlling gene expression via, 89–91
enhancers in, 91–93
in *Luzula* chromosomes, 4
methylation and, 57
mutagenesis and, 78
of plants, ix–x, xi
in reassociation experiments, 55, 56
of rice, 79

sequencing of, 23, 39
single-nucleotide polymorphisms in, 152
DNA fingerprinting, physical mapping and, 51–53
DNA markers, 48, 149–152. *See also* Marker-assisted selection (MAS)
DNA microarrays, 37–38. *See also* Microarray technology
advantages and disadvantages of, 119–120
error estimating in, 112–113
types of, 110
DNA polymerase, in cDNA cloning, 32
DNA-Protein Search program, 179
DNA values, table of selected, 2
DNA variation
origin of, 7–16
in plants, 1–3
Dof gene family, 98
Double-stranded RNA (dsRNA), 86, 87
Doubling, of chromosomes, 13–14, 14–15
Downstream element (DST), 99
Downstream processing, 24
Drought stress, 131, 142–144
Ds transposon, 84
Dynamic data, 169

EDS1 gene, 136
Electrospray ionization (ESI), 39
Enhancement, intron-mediated, 95
Enhancer elements, effect on gene expression, 91–93
Enhancers, defined, 91–93
Enhancer traps, 82, 84–85
Environment, plant interactions with, 131–144
Enzymes
gene families and, 69
methylation-sensitive, 57
of plants, ix
Error estimating, in microarray experiments, 112–113
Escherichia coli, 29, 57, 61
EST assembly, in informatics, 172–173
EST clustering, informatics for, 42–43, 179–180
EST Databases, 63
EST libraries, in informatics, 172–173
EST markers, 52–53
EST numbers, table of, 74
Estradiol-inducible ER-C1 system, 100
Estradiol-inducible XVE system, 100
Ethanol-inducible AlcR system, 101
Ethics, plant genomics and, 189–198
Euchromatin, 57
Eukaryotic promoters, conserved sequences in, 95–96
European Union (EU), moratorium on GMOs by, 192, 193–196
Evolution, chromosomal gene distribution and, 16–19
Expressed sequence tags (ESTs), 30. *See also* EST entries
in cDNA cloning, 32–33
cDNAs versus, 77
in gene expression analysis, 110, 113–114, 119
in gene identification, 70–73, 74
generating, 71–73, 74
genome fractionation via, 54, 55
RFLPs and, 149–150
software for clustering, 42–43, 179–180
Expression data, 179
Expression profiling, 108, 109–120
in pathogen resistance studies, 137–138
Expression vectors, 25
Extracellular LRR proteins, 133, 134, 136, 137
Extraction, isolating source DNA via, 28

FASTA file format, 42
FASTA system, 178
Fingerprint contigs (FPC) software, 51–53
Fingerprinting
BAC end sequencing and, 59–60
physical mapping and, 51–53
5′ sequences, in gene expression, 93
Flax
BLAST® sequences for, 183–185
disease resistance in, 132–133
Miropeats analysis of, 177

Flax polyubiquitin gene, BLAST® sequences for, 183–185
Floral Genome Project, xi
Flowering plants, origin of, xi
Fluorescent dyes, in protein separation, 124–125
FootPrinter program, 182
Fractionation, of genomes, 53–58, 61–62
Fritillaria assyriaca, genome size of, 2
Full-length cDNAs, 73–75, 75–77
Fungi, mycorrhizal, 138–139
fw2-2 locus, in tomato, 156–160

Gateway™ Technology, cloning via, 25–27
GenBank, 170, 173
GenBank dbEST database, 72–73
  plant EST numbers in (table), 74
GenBank Nr (nonredundant) database, 63
Gene enrichment strategy, 48–49
Gene expression. *See also* Expressed sequence tags (ESTs)
  chemically regulated, 99, 100–101
  control of, 89–103
  enhancer elements and, 91–93
  5′ sequences in, 93
  introns in, 94–95
  promoters for, 91, 92
  proteomics and, 120
  3′ sequences in, 93–94
  trans-acting factors controlling, 96–98
  translational control in, 103
Gene families, 69–70
  in *Arabidopsis* (table), 97–98
Gene-for-gene interaction, 132–134
Gene indices, 179, 180
Gene knockouts, 80
  Arabidopsis, 81
Gene linkage, 18–19
Gene number, counting, 9–10
Gene Ontology (GO) project, 169
Genes. *See also* Genomes
  avirulence, 133–135
  bioinformatics and, 167
  complex traits and, 147–164
  conditioning QTLs via, 156
  defined, 69
  discovering, 69–87
  disease resistance, 133–138
  drought stress in plants and, 142–144
  encoding ribosomal RNA, 13
  identifying, 69–70, 70–73, 78–87
  informatics tools for finding, 180
  phenotypes and, 147–149
GeneSeqer program, 180, 185
Gene silencing, 16, 24, 80, 86–87
Gene space, 53–54, 65–66
Genetically modified organisms (GMOs)
  bioethics of, 190–191
  future of, 197
Genetic maps, in genome sequencing, 50–51
Gene traps, 82
Genome browsers, 182–183
Genome collinearity, 17
Genome contraction, hypothetical effects of, 15
Genome expansion, hypothetical effects of, 15
Genomes. *See also* Genes; Genomics
  annotation of, 77–78, 79
  automated sequencing strategies for, 38
  cellular processes and, 107–109
  creating complete sequences of, 47–66
  fractionation of, 24–25, 53–58, 61–62
  functional data from, 23–43
  informatics in analyses of, 171
  multiple, 16–19
  sequencing strategies for, 58–61
  structure of plant, 1–20
  tools for investigating, 23–24
Genome size
  of plants, 1–3
  processes affecting, 13–16
Genomics. *See also* Genes; Genomes
  bioethics of, 189–198
  comparative, 182–183
  functional, 107–129
  goals of, 69
Genomic sequencing, informatics tools for, 180
Global community, data sharing within, 41

Gramene database, 42, 168, 170, 183
Grasses
  gene linkage in, 18–19
  genome size increases in, 15
GST (glutathione-*S*-transferase) genes, 8, 9
GUS (glucuronidase) activity, 82, 83, 84–85, 95

Haplopappus gracilis, chromosome number of, 3
HarvEST program, 179–180
HB gene family, 98
Heat shock cognate 80 (HSC80) gene, 102
Heterochromatin, 20
High-throughput TILLING, 84
Histone core, in chromatin, 102
HMM profiling, 178
Holocentric chromosomes, 4
Human genome, EST collections for, 54
Human Genome browser, 177
Hybridization. *See also* Yeast two-hybrid system
  in BAC screening, 30, 31, 32
  in microarray experiments, 110–113
  RFLPs and, 149–150
Hypermethylation, 57
Hypersensitive response, 133, 134
Hypomethylation, 57

Immobilized pH gradients (IPGs), 124
Ineffective nodulating mutants, of legumes, 139
Informatics tools, 24, 39–43, 170–180. *See also* Bioinformatics; Software
Insertional mutagenesis, 24, 80–82, 83, 84–85
  in pathogen resistance studies, 138
International Mapping Consortium, 63
International Rice Genome Sequencing Project (IRGSP), 48, 49
International Sequencing Consortium, 63
Intron-mediated enhancement (IME), 95
Introns, in gene expression, 94–95
Isoelectric focusing (IEF), 124

Jasmonic acid/ethylene-mediated responses, 136–137
Joint Genome Institute, 65

Kinetochore activity, in *Luzula*, 4
Knockouts, 80
  Arabidopsis, 81

Large-insert vectors, in cloning, 27
Legume Information System, 168, 170
Legumes
  mutants of, 139–141
  nitrogen-fixing bacteria and, 138–139
  as sequencing targets, 65
Leucine-rich repeat (LRR) genes, 133–134, 136, 137
Libraries
  BAC, 28
  Cot fractionation-based, 62
  EST, 172–173
  full-length cDNA, 73, 75
  in identifying QTLs, 160
  large-insert, 27
  methylation restriction, 62
  methyl filtration, 61–62
  shotgun, 61–62
  subtraction, 33–36
Light, plants and, xii
Living modified organisms (LMOs), environmental release of, 191, 193–196
Loci. *See* Quantitative trait loci (QTLs)
Long terminal repeats (LTRs), 10
Lotus japonicus, 139, 140
  as sequencing target, 65
Lotus tenuis, chromosome variation in, 4, 6–7
Low-copy sequences, DNA variation via, 7–10
LR reaction, 25–27
Luzula, chromosome variation in, 4
Luzula elegans, chromosome variation in, 4, 5
Luzula pilosa, chromosome variation in, 4, 5, 6
Lycopersicon, QTLs in, 156–160
Lycopersicon esculentum, 157, 158–159
Lycopersicon pennellii, 158–159

Lycopersicon pimpinellifolium, 157
Lynx technology, 119

MADS gene family, 98
Maize
  gene linkage in, 18–19
  genome sequencing of, 53
  molecular maps for, 149
  retrotransposons in, 10–11
  as sequencing target, 65
  transposable elements in, 57–58
  transposon mutagenesis in, 82, 84
MaizeDB database, 168, 170, 183
Mannopine synthase (MAS) promoter, 93, 94
Map-based cloning, 48
Map-based genome sequencing, 49, 50
Mapping, of QTLs, 152–160. *See also* Genetic maps; Maps; Physical mapping
Maps, molecular, 148, 149
Marker-assisted selection (MAS), 148–149, 160–164
Massively parallel signature sequencing (MPSS™), 110, 113, 114, 117–119
  advantages and disadvantages of, 119–120
Mass spectrometry (MS), 39, 40
  in protein identification, 125
Matrix-assisted laser desorption/ionization (MALDI), 39
Matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF), 121
Matrix attachment regions (MARs), 102
Medicago truncatula, 139, 140
  as sequencing target, 65
Megablast program, 179
Messenger RNA (mRNA), 75, 120
  in cDNA cloning, 32–33
  in differential display technology, 36–37
  MPSS and, 118
  SAGE and, 116
  stability of, 99
  subtraction libraries and, 33
Metabolites
  cellular processes and, 107–109
  synthesized by plants, xii–xiii
Metabolomes, 107–109
Metabolomics, 109
Metaphase, in *Luzula* species, 5, 6
Methoxyfenozide-inducible GVCfEcR system, 101
Methoxyfenozide-inducible GVOnEcR system, 101
Methylation, 57
Methylation restriction libraries, 62
Methylcytosine (Methyl-C), 57
Methyl filtration
  genome fractionation via, 57
  libraries, 61–62
Microarray hybridization, 48
Microarray technology, 24, 37–38, 96. *See also* DNA microarrays
  in expression profiling, 109–110
  in plant stress studies, 132, 142–143
Microsatellites, 149, 151
Miller, Webb, 182
Miniature inverted repeat transposable elements (MITES), 11
Minimal information associated with a functional genomics experiment (MIAFGE), 42
Minimum tiling path (MTP) sequencing, 59, 60, 61
Miropeats program, 174, 177, 178
Molecular maps, 148, 149
Monocots, intron-mediated enhancement in, 95
Motif finding, 178
MultiPipMaker program, 182–183
Multiple genomes, consequences of, 16–19
Musa, 32
Musa acuminata, 32
Musa balbisiana, 32
Mutagenesis
  gene identification via, 78–87
  insertional, 80–82, 83, 84–85, 138
  via RNA interference, 80, 86–87
  via TILLING, 80, 82–85
  transposon, 82, 84
Mutagens, 70

Mutants, of legumes, 139–141. *See also* Mutagenesis
MYB gene family, 97
Mycorrhizal fungi, 138–139

NAC gene family, 97
National Biosafety and Biotechnology Strategy Plan (Zambia), 194
National Center for Biotechnology Information (NCBI), 168, 173
  databases available from (table), 176
National Plant Genome Initiative, xiv, 66
NDR1 gene, 135–136
Near isogenic lines (NILs), in QTL identification, 152–155, 160
Nitrogen-fixing bacteria, 138–139
Nonnodulating mutants, of legumes, 139
Nonreciprocal translocations, 14
Normalization, of cDNA libraries, 34
Northern blot, in expression profiling, 109–110
Nuclear DNA, in plants, 1–3
Nuclear isolation, isolating source DNA via, 28
Nucleic acid searches, 183–185
Nucleolar dominance, 16
Nucleosome, 102
Nucleotide sequences, BLAST® and, 175, 176
Nucleotide-binding leucine-rich repeat (NBS-LRR) genes, 133–134, 136

Oligonucleotides. *See also* Overgo entries
  as BAC library probes, 30, 31
  in microarrays, 110, 112
1C values, 1–3
Open reading frame X (ORFX), in identifying QTLs, 158–160
Organellar "contamination," 61
Organ-specific promoters, 148
Orion Genomics, 61
Orthologs, 18
"Overgo oligos," 30, 31
  with BAC contigs, 58
Overgo probes, 58

PaCE clustering tool, 179
PAD4 gene, 136
Paracel TranscriptAssembler™, 179
Paralogs, 18
PAT1:β-glucuronidase (PAT1:GUS) fusion, 95
Pathogens, xi–xii
  plant resistance to, 133–138
PBS2 gene, 136
PCR, real-time, 129
PCR amplification, 151
PCYPAC2N vector, 29
Peptides, sequencing of, 38–39, 40, 121, 122
Peronospora parasitica, 135
Pests, 131
PHD file format, 42
Phenotypes, 19
  genes and, 147–149
Phenotypic screening, in marker-assisted selection, 160–164
Phrap software, 42–43
Phred software, 42–43
Physical mapping, 51–53, 64–65
Physical maps, in genome sequencing, 49, 50–51
Pinus pinaster, DNA content of, 2–3
Pita protein, 137
Plant breeding, bioethics of, 189–190
Plant databases, critique of, 41–42. *See also* Databases
PlantGDB database, 179, 180
Plant Genome Research Program, 42, 132
Plant genomes. *See also* Genomes; Plant genomics
  creating complete sequences of, 47–66
  structure of, 1–20
Plant genomics. *See also* Genomics; Plant genomes
  bioethics of, 189–198
  informatics for, 171
  rationale for, ix–xiv
Plant roots, interactions of, ix, xi
Plants
  abiotic interactions of, 141–144
  annotation of genomes of, 77–78
  for BAC libraries, 32
  biotic interactions of, 132–141

chromosome variation in, 3–7
complex traits of, 147–164
controlling gene expression in, 89–103
differences between other organisms and, ix–x
disease resistance of, 131, 132–138
environmental interactions of, 131–144
EST numbers of selected (table), 74
features of, ix
gene spaces of, 66
identifying gene families in, 69–70
proteomics of, 124
recognition of pathogens by, 135
sequence databases for, 65
stress among, 131–132
symbiosis among, 132, 138–141
PlantsP database, 181
PlantsP Kinase Classification (PPC), 181
Plasmid-based vectors, in cloning, 25–27
Plasmids, 126, 127
Polyacrylamide gel electrophoresis (PAGE), in protein separation, 124
Polymorphisms, genetic maps and, 51. *See also* Amplified fragment length polymorphisms (AFLPs); Random amplified polymorphic DNAs (RAPDs); Single-nucleotide polymorphisms (SNPs)
Polyploidy, 1
consequences of, 16–19
genome size and, 13–14
Poplar, as sequencing target, 65
"Prey" plasmids, 126, 127
Promoters
conserved sequences in, 95–96
eukaryotic, 95–96
for gene expression, 91, 92
MAS, 93, 94
organ- and tissue-specific, 148
Prosite, 178, 180–181
Protein homology searches, 183–185
Protein identification, in proteomics experiments, 125
Protein isolation, in proteomics experiments, 123–124
Protein-protein interactions, in proteomics experiments, 125–126
Proteins. *See also* Proteomes; Proteomics
cellular processes and, 107–109
drought stress in plants and, 142–143
informatics tools for, 180–181
in proteomics, 120–128
recognition of pathogen, 135
sequencing of, 38–39, 40
Protein separation, in proteomics experiments, 124–125
Protein sequences, BLAST® and, 175, 176
Protein tags, in proteomics, 126–128
Proteomes. *See also* Protein entries; Proteomics
analysis of, 39, 40
cellular processes and, 107–109
Proteomics, 109, 120–128. *See also* Proteomes
informatics tools for, 180–181
in pathogen resistance studies, 138
in plant stress studies, 132
transgenics in, 126, 128
yeast two-hybrid systems in, 126, 127
Public data archives, 170
Public perception, of genetically modified organisms, 193–196
Public policy, toward genetically modified organisms, 196–197
Pumilio-like protein, 8, 9

Quantitative trait loci (QTLs), 48, 148
genes conditioning, 156
identification of, 152–160
in tomato, 156–160

Random amplified polymorphic DNAs (RAPDs), 149, 151
Reactive oxygen species (ROS), 141
Real-time PCR, 129
Reassociation kinetics, genome fractionation via, 54–56
Recognition, of plant pathogens, 135
Recombinant DNA, 23
Recombination, genetic maps and, 50–51

Regulation, of genetically modified organisms, 191, 193–196
Relational databases, 167
Resistance (R) genes, 133–138
Restriction enzymes, methylation-sensitive, 57
Restriction fragment length polymorphisms (RFLPs), 149–150
Retrotransposons, 10–11
  in genome size increase, 15
Reverse transcriptase, in cDNA cloning, 32
Reverse transcription PCR (RT-PCR), 173
R genes. *See* Resistance (R) genes
Rhizobia, as sequencing target, 65
Rhizobium, 139–140
Ribosomal RNA (rRNA), 11, 13, 16
Rice
  annotated DNA sequence of, 79
  centromeres of, 12–13
  drought-stress proteins in, 143
  genome sequencing of, xiii, 48, 53
  genome variation in, 7–8, 10
  GUS activity in, 82, 83
RNA, 129. *See also* Double-stranded RNA (dsRNA); Messenger RNA (mRNA); Ribosomal RNA (rRNA)
  abnormal, 70
  in cells, 107–108
  in controlling gene expression, 89–91
  counting molecules of, 113
  in expression profiling, 109–110, 111–113
  in generating ESTs, 71, 73
  mutagenesis and, 78–80
  in proteomics experiments, 121
RNA interference (RNAi), 24, 80, 86–87
RNA polymerase, 90
RNA polymerase III products, 15
Rosa, genome size of, 2
Royal Botanic Gardens, 2

Safener-inducible In2-2 system, 101
Safety, of genetically modified organisms, 190, 191–192, 193–196
SAGE tags, 114–115, 116
Salicylic acid-mediated responses (SAR), 136–137
SCF file format, 42
Sedum suaveolens, chromosome number of, 3
Segregating populations, in QTL identification, 152–154, 160
Sequence assembly, software for, 171–172
Sequence data, gene identification via, 70–73
Sequence matching, BLAST® and, 174, 175, 176, 177, 178
Sequence quality, 42–43
Sequence similarity, 174
Sequence-tagged connectors (STC) approach, BAC end sequencing and, 60
Sequence-tagged sites (STSs), 151
Sequencing, 70
Sequencing strategies, 47–66, 58–61
  automated, 38
  data processing for, 63
  most efficient, 63–64
  targets for, 64–66
Serial analysis of gene expression (SAGE), 110, 113, 114–116, 119
  advantages and disadvantages of, 119–120
Shotgun libraries, 61–62
Shotgun sequence analysis, 48–49, 50
  data processing for, 63
  of poplar genome, 65
Signal transduction networks, in plant disease resistance, 135–138
Silencing, 80, 86–87
  of genes, 16
Simple sequence repeats (SSRs), 149, 151
Single-nucleotide polymorphisms (SNPs), 43, 149, 152
Small auxin-up RNA (*SAUR*), 99
Software. *See also* Bioinformatics; Informatics tools
  database-independent, 41
  sequence quality from, 42–43
Soybean, molecular maps for, 149
Spatial regulation, 89

Spliceosome formation, 95
Splicing, alternative, 112
Spontaneous doubling, of chromosomes, 13–15
Spontaneous nodulation mutants, 139
Staining protocols, in protein separation, 124–125
Static data, 169
Stress, among plants, 131–132, 141–144
Subtraction libraries, 33–36
Suppressive subtracted hybridization (SSH), 33, 34–35
Symbiosis, 132, 138–141
Syngenta, 48, 49
Synteny, 16–18, 50, 70, 78
  informatics tools for, 183
  testing of, 30–32

Tandemly repeated sequences, DNA variation via, 11–13
Tandem MS, in protein identification, 125
Targeting induced local lesions in genomes (TILLING) methodology, 80, 82–86
T-DNA insertion, 80–82, 84
Tebufenozide-inducible GVHvEcR system, 101
Technology, improvements in, 23. *See also* Biotechnology; Microarray technology
Temporal regulation, 89
Tentative contigs (TCs), 71, 73
  in informatics, 172–173, 179
Tentative unique genes (TUGs), 71, 73, 179, 180
teosinte branched1 gene, 160
Tetracycline-inactivatable tTA system, 100
Tetracycline-inducible TetR system, 100
The *Arabidopsis* Information Resource (TAIR), 168, 170
3′ sequences, in gene expression, 93–94
TIGR Gene index, 179, 180
TIGR Web site, informatics at, 179, 180
TIR-NBS-LRR gene family, 133, 134, 136
Tissue-specific promoters, 148
Toll and Interleukin-1 receptor (TIR) genes, 133
Tomato, *fw2-2* locus in, 156–160
Torrey Mesa Research Institute, 48, 49
Trade, in genetically modified organisms, 191, 192
trans-acting factors, 96–98
Transcription factors, gene expression and, 96–98
Transcriptomes, 107–109
Transgenic crops, bioethics of, 190–191
Transgenics, 77
  in proteomics, 126, 128
Translational control, in gene expression, 103
Transposon mutagenesis, 77, 82, 84
Transposons, 10–11, 20, 70
  in insertional mutagenesis, 80, 82, 84
Transposon tags, genome fractionation via, 57–58
Tropinone reductase genes, 8, 9
Two-dimensional gel electrophoresis (2-DE), 39, 40
  in protein separation, 124–125
Type III secretion system (TTSS), 135

UNEP International Guidelines, 193–194
Unequal recombination, 14
Unigene sets, in informatics, 172–173. *See also* Tentative unique genes (TUGs)
United States, support for plant genomics in, xiv

Vectors
  large-insert, 27
  plasmid-based, 25–27
Vicia faba, chromosome variation in, 6–7

Water, drought stress in plants and, 142
Whole genome approach, to plant genomics, xiii
Whole genome shotgun (WGS) approach, 63–64

World Trade Organization (WTO), 192
WRKY domain, 134
WRKY gene family, 98

YAC libraries, 160
Yeast artificial chromosomes (YACs)
  in identifying QTLs, 158, 160
  large-insert vectors and, 27
Yeast two-hybrid systems, in proteomics, 126, 127

Zambia, genetically modified organisms from, 194
ZC_2H_2 gene family, 98
Zinnia, x
ZmDB database, 42, 168, 170